Unclassified

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)*
Bolt Beranek and Newman Inc.
50 Moulton Street
Cambridge, Massachusetts 02138

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3. REPORT TITLE
COMMAND AND CONTROL RELATED COMPUTER TECHNOLOGY

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*
Quarterly Progress Report, 1 Dec 74 to 28 Feb 75

5. AUTHOR(S) *(First name, middle initial, last name)*
Jerry Burchfiel, R. Viswanathan, Raymond S. Nickerson

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
March 1975	96	10

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
MDA903-75-C-0180	
b. PROJECT NO. ARPA on 2935	BBN Report No. 3064
c.	9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*
d.	

10. DISTRIBUTION STATEMENT
Distribution of this document is unlimited. It may be released to the Clearinghouse, Department of Commerce for sale to the general public.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
This research was supported by the Advance Research Projects Agency under ARPA Order No. 2935	

13. ABSTRACT

This document describes: 1) work performed in the design and development of a Packet Radio Network which supports a variety of command and control requirements for mobile digital communications. 2) new developments in speech compression to support low bandwidth (<2 kilobaud) digital speech for transmission over packet-switched networks. 3) an evaluation of the quality of vocoded speech to permit the direct comparison of different systems as well as improve the performance of any given system.

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

Unclassified

Security Classification

A-31408

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
packet radio						
computer communications						
packet switched networks						
cross-network debugging						
ELF						
BCPL						
speech compression						
vocoder						
linear prediction						
parameter quantization						
optimal parameter interpolation						
real time signal processing						

TABLE OF CONTENTS

I. COMMAND AND CONTROL STUDIES

During this quarter several Naval installations in the San Diego area were visited as follows.

Naval Air Station, Miramar
- Commander, Air Wing 2
- Air Combat Maneuvering Range
- F14 training and simulator facilities

Naval Air Station, North Island
- Commander VS-22
- S3A training and simulator facilities
- S3A operational briefing facilities

Naval Electronics Laboratory Center
- Technical Director
- Computer Science Department
- NTDS development facilities

Discussions centered on the current state of military tactical computing and problem areas. In addition the level of interest and suitability of these facilities for a tour during the ARPA/IPTO Principal Investigators conference was explored. NELC did arrange and host the tour on March 14.

II. PACKET RADIO NETWORK

A. Meetings

Three major meetings were held this quarter to discuss Packet Radio Network system design issues.

The first was held November 18 and 19 at the Stanford Research Institute. The result of this meeting was a specification of the fall 1975 area test to be conducted at SRI, along with preliminary scheduling estimates by all contractors.

The second meeting was held at Network Analysis Corporation on December 19, 1974 between NAC and BBN to begin exploring protocol issues. The result of this meeting was a set of protocol issues to be resolved in a wider forum.

The third meeting was held at ARPA, March 6 and 7, with all contractors attending to attempt a specification of the protocols to be used in the Packet RAdio Network. Preliminary specifications were reached for five protocols. Radio Control, Channel Access, Source-to-Destination Transport, Station-to-Terminal, and Kahn-Cerf Internet. Each of these protocols is now undergoing continuing refinement.

B. Publications

A major milestone was met this quarter with publication of "Functions and Structure of a Packet Radio Station," Packet Radio Temporary Note #125. This met the March 15 milestone for specification of the Packet Radio Station hardware, operating system, and applications programming environment. This same paper will be presented at the Packet Radio session of the AFIPS 1975 Joint Computer Conference.

To aid in details of joint project management planning, we also generated and negotiated with Collins Radio a detailed specification of our acceptance test for the Packet RAdio Digital units to be delivered to BBN by Collins. Delivery of these units has unfortunately been delayed by the formal procedures required to move a piece of government equipment from one contractor to another.

In addition, we published the following notes on Packet Radio System design issues:

Tomlinson, R.S., Selecting Sequence Numbers, August 1974.

Tomlinson, R.S., Packet Radio System Design Issues, PRTN 122, August 1974.

Burchfiel, J.D., Packet Radio System Capabilities, PRTN 123, September 1974.

Tomlinson, R.S., Proposed PRN Protocols, PRTN 124, October 1974.

Burchfiel, J.D., Functions and Structure of a Packet Radio Station, PRTN 125, December 1974.

Tomlinson, R.S., Point-to-Point Routing in the Packet Radio Network, PRTN 126, January 1975.

C. Cross-net Debugger

1. Description

The distribution of Packet Radio Units throughout an area of many square miles necessitates a new approach to debugging. Indeed, the debugging of IMP computers in the ARPA Network has necessitated a first step in this direction. Each IMP has its own debugger program resident in IMP memory. A network programmer communicates with this debugger character by character over the network. The characters are assembled into commands within the IMP, the command performed, and any response prepared as a sequence of characters. This response characters are then transmitted back over the network to the programmer's terminal. Remote placement of the terminal, as occurs in this case, is the first step in debugger evolution for the Packet Radio Network.

The second evolutionary step is to place the major computational portion of the debugger in a separate machine. This step has been taken in the form of X-NET, a cross-network debugger in which a powerful "controlling" machine (PDP-10 computer) is used to perform the more complex debugging tasks. The program being debugged resides in a "target" machine (a relatively small PDP-11 computer). Both machines are connected to the ARPA Network; thus they have a means for exchange of information. A small, simple debugging process runs in the target machine in parallel with the program being debugged. The complex part of X-NET in the controlling machine sends simple commands over the network to the simple process in the target machine, where each is acted upon and a

reply returned. The controlling machine then prints any relevant data regarding the interaction for the programmer to see. This system of debugging has been implemented as has been shown to function usably (see section 2, below).

Looking forward to the future, the third and ultimate step will be a debugger similar to X-NET, but using radio transmission links instead of or in addition to the ARPA Network, to debug Packet Radio Units as target machines. X-NET is an important tool in developing and debugging Packet Radio Station programs, as the station is a PDP-11 computer. Experience gained has the additional benefit of guiding future development of the "cross-radio" debugger.

2. Demonstration

On February 20 the X-NET debugger was demonstrated to members of the Stanford Research Institute, via terminal linking on the ARPA Network. We at BBN typed X-NET commands to perform the following actions on a target machine (host 205, a PDP-11 in a room nearby):

(a) create a process under the target machine operating system, ELF;

(b) load a program from a disk file on the controlling machine into the memory of the target machine;

(c) start the program running;

(d) halt the program;

(e) examine contents of locations in the target machine memory, and deposit new values;

(f) set a breakpoint in the program, receive notification that the breakpoint has been reached, and proceed from the breakpoint;

(g) search locations in the program for a particular value;

(h) dump the program back into a disk file on the controlling machine; and finally

(i) delete the process from its existence under ELF.

As we performed these manipulations, we demonstrated various ways of typing in and typing out values (as numbers, as characters, etc.); that multiply proceeded breakpoints have the proceed count maintained in the target machine for speed; and that debugger operations can be performed simultaneously with execution of the program being debugged. In addition, we demonstrated a statistics feature of the X-NET debugger, which is explained in further detail below. The text which we typed and the replies typed by X-NET were both echoed on the monitoring terminal at SRI.

3. Statistics

The X-NET debugger keeps a record of its own use of the ARPA Network (both in terms of number of messages and in terms of total amount of information), and a record of response time which the network has exhibited during the debugging session. This permits analysis of overhead experienced in this form of debugging, and adjustment of tradeoffs in the operation of X-NET to minimize this overhead. The delays experienced are usually not inconveniencing on a human scale. The lump averages and standard deviations, however, are inadequate for a proper understanding of certain factors. One factor is how expensive various commands are, in terms of delay; do certain commands cause particularly large delay? Another factor is variation in response time due to vagaries of the controlling machine's operating system (TENEX) and the network alone (without regard to time X-NET spends computing). For these reasons, a more complete statistics function is anticipated in the near future. This, plus some features to increase the ease of using X-NET, are required before it is a fully functional system.

D. ELF Development

Since the Packet Radio Station will run on a PDP-11 computer under the ELF operating system, we have devoted considerable effort to getting the ELF system running on our PDP-11, and in making changes to the system to facilitate the work on the station software. The major portion of this effort, unfortunately, was locating up-to-date and consistent source files for the ELF system. Once this was done, the system changes were relatively easy to implement.

The changes to ELF were made to facilitate the implementation of the cross-net debugger server which will be used to debug the station software. A breakpoint handler was added to the ELF system. This handler notifies the debugger server process whenever a user process executes the BPT (BreakPoint Trap) instruction. Other changes were added to allow the debugger server process to manipulate a user processes registers, and to stop a user process from running (Freeze it) in such a way that the creator of the process is not notified. The process register manipulation primitive in ELF will need further changes to prevent the manipulation of registers while a process is executing in KERNEL mode (i.e. during system primitives) and to manipulate the saved user mode registers instead. This is necessary both to protect the system from the user, and to avoid confusing the user with information which is of no use to him.

An initial version of the debugger server process has been implemented and all functions it should perform have been checked out. Some minor changes are needed to protect the debugger server from users who try to manipulate it. Also the server needs to be expanded to support more than one debugging session at once. The framework for this already exists and all that is needed to make it functional is the addition of a few interlocks on sensitive tables and the expansion of the tables, (which currently will only hold information about one debugging session).

E. BCPL Runtime Support

The BCPL language has been selected for implementation of Packet Radio Station software. This software will run under the ELF operating system for the PDP-11 computer. BCPL, like any higher-level language, provides various services and canned routines for the convenience of the user. Many of these services and routines involve input from, and output to, peripheral devices attached to the PDP-11. When there is no operating system present, the BCPL routines may access these peripheral devices directly. Under an operating system, however, the devices must usually be accessed through calls to that system; otherwise the system itself would be disrupted. Hence, these routines, referred to as the BCPL "RUNTIME" routines, had to be modified. The modifications necessary proved more extensive than first estimated, principally due to major revision of information input and output control structure.

Briefly, the changes consisted of:

(1) Modifying the "RUNTIME" routines to use the ELF operating system's "freeze process" call to gracefully terminate execution of the user's program when its computation completes.

(2) Adding two routines to the menu the BCPL user has available, namely "ELFCALL", which allows the user direct access to any and all of the services which the ELF operating system provides, and "CREATE", which creates allocations within ELF for a new process (Program). "CREATE" is also accessible to the BCPL user as a particular instance of "ELFCALL", but it was redundantly added to allow the user a simpler calling sequence.

(3) Rewriting the BCPL routines which handle ARPA Network messages. These routines used to manipulate the PDP-11's interface to the IMP directly. Since this is incompatible with the ELF operating system's simultaneous use of the IMP interface, the routines now use ELF conventions to do ARPA Network input and output through ELF.

(4) Revising other BCPL input and output routines to use ELF conventions instead of direct access to peripheral devices. The motivation for these change was analogous to the rewriting of

the IMP interface routines mentioned above.

(4) Especial modification to a particular BCPL routine. "AnyInput". This required a major rethinking of the control structure of information input, and is discussed separately at greater length below.

(6) Modifying or creating files of declarations and definitions, to support the changes mentioned in the five items above.

One of the routines provided by the BCPL language for the convenience of the user is called "AnyInput". It allows the user's program to test whether there is any information (on a specified input stream) which is available but not yet actually read by the user's program. It simply tests whether the specified input queue is empty, and return a "yes, there's input data" or "no, there's none" to the user's program. The important aspect is that "AnyInput" merely tests the emptiness of the queue; it does not wait for data if none is available; hence it does not suspend execution of the user's program. Unfortunately, there is no direct way to accomplish this function under the ELF operating system. The only call to ELF which tests an input queue also waits, with the user's program not executing, until data has arrived. The solution found was to have the "AnyInput" routine send a special message to itself, in a format distinguishable from normal input data. The ELF operating system places this message at the end of any other data already queued. "AnyInput" then reads one data item; if it is this special message, there was no data available and "AnyInput" returns the answer, "no" to the user's program. If, however, a legitimate data item is read, then "AnyInput" must save this data in a special place. It also must set a flag so the routines which service actual reading commands from the user's program will give this data to the

program first. "AnyInput" may then return the answer, "yes" to the user's program. Of course, "AnyInput" checks its own flag before sending itself the special message -- if a data item is already immediately, "yes!" The moral to the "AnyInput" situation is that input and output conventions of operating systems and of high-level languages, which are often specified in great detail, are sometimes difficult to wed.

One further change was made to BCPL output routines. The "CREATE" call mentioned above is used to create a special process to do double buffering of output data. By employing two buffers, the user's program can be filling one while the other, previously filled with data, is being handed to ELF and sent out to the peripheral device in question. This old computer technique allows overlapping of the two operations, assembling output data and transmitting the data. Use of a special process to implement this shuffling of buffers achieves efficient operation.

TABLE OF CONTENTS

I. SPEECH COMPRESSION

In our speech compression project, we continue to improve the quality of speech transmitted at low bit rates. The development of a new optimal linear interpolation scheme for receiver parameters has been the major result of our research in the last quarter. This scheme requires the transmission of an extra coefficient per data frame, which carries information about interpolation. The consequent increase in transmission rate is only slight. However, experiments using both variable frame-rate and constant frame-rate transmission schemes have shown this optimal linear interpolation scheme to be superior to simple linear interpolation, especially during rapid transitions in the speech signal. Next, we proposed a new pitch quantization procedure which makes the best use of all the quantization levels. Also in the last quarter, the central component of our proposed speech processing system, the SPS-41 computer, was delivered. We have begun work in organizing the PDP-11/SPS-41 system.

A. Optimal Linear Interpolation (OLI)

In low bit-rate linear predictive speech compression systems, the process of parameter interpolation at the receiver helps in smoothing the roughness in the synthesized speech which is normally associated with infrequent parameter updating. Simple linear interpolation (SLI) has been used almost exclusively in these systems. In an earlier study we found that the spectral error due to interpolation was much larger than the error due to quantization (BBN Report No. 2976, p. 96). This result suggests that better parameter interpolation approaches than the simple linear scheme should be investigated. With this motivation, we have developed an optimal linear interpolation (OLI) scheme that requires the transmission of an extra parameter per data frame, α: $0 \leq \alpha \leq 1$. The value of α is determined as that point along the line used for linear interpolation which is closest (in the mean square sense) to the point determined by the actual parameter values at the instance where interpolation is desired. The transmission of α requires 50-150 bits/sec, depending on the frame rate and the number of bits used for quantizing α.

Below we present theoretical and experimental results that we obtained with the new interpolation scheme. Theoretical results show that in the space of parameter vectors, the OLI scheme corresponds to an orthogonal projection of the actual parameter vector at the interpolation point onto the line passing through the two parameter vectors that are used in the interpolation. We present reasons for our choice of log area ratios (LARs) for use in

OLI. Several ways of using the OLI scheme with a variable frame-rate transmission system are also given. Experimental results show that the OLI scheme improves speech quality relative to the SLI scheme, especially during rapid transitions in the speech signal. In addition to informal listening tests, we have investigated the waveforms and spectrograms of synthesized speech with OLI, and the time history of the spectral error.

1. Derivation of Optimal Linear Interpolation

Let $\underline{g}_1$, $\underline{g}_2$ and $\underline{g}_3$ denote p-dimensional parameter vectors* for frames n, n+1 and n+2. It is assumed that the transmitter transmits $\underline{g}_1$ and $\underline{g}_3$, and the receiver performs some form of interpolation over the received parameters $\underline{g}_1$ and $\underline{g}_3$ to generate an approximation to $\underline{g}_2$. The line joining $\underline{g}_1$ and $\underline{g}_3$ is described, in the p-dimensional parameter space, by the expression

$$\underline{g} = (1-\alpha)\ \underline{g}_1 + \alpha\ \underline{g}_3 \quad . \tag{1}$$

For SLI, if the frames are equally spaced then $\alpha = 1/2$, i.e., $\underline{g}_2$ is approximated by the arithmetic mean $(\underline{g}_1 + \underline{g}_3)/2$. In OLI, we define the interpolation to be optimal if α is selected so as to minimize

---

*By parameters, we mean those that characterize the p-th order linear predictor. Gain and pitch are excluded. For the purposes of this section, we need not specify which parameters are used for interpolation. The choice of interpolation parameters is the topic of the next section.

the total-squared error

$$E = (\underline{g} - \underline{g}_2)^T (\underline{g} - \underline{g}_2) = \sum_{i=1}^{p} (g_i - g_{2i})^2 , \qquad (2)$$

where superscript T denotes transpose, and g_i and g_{2i} are the i-th components of the vectors $\underline{g}$ and $\underline{g}_2$ respectively. By equating $(\delta E/\delta\alpha)$ to 0, we find that the optimal value of α is given by

$$\alpha^* = (\underline{g}_2 - \underline{g}_1)^T (\underline{g}_3 - \underline{g}_1) \; / \; (\underline{g}_3 - \underline{g}_1)^T (\underline{g}_3 - \underline{g}_1) \; . \qquad (3)$$

Since we consider below only the optimal value of α, we shall omit the superscript * for notational convenience. It can be easily verified that the second derivative of E with respect to α evaluated at the optimum is nonnegative. (It is zero if and only if $\underline{g}_1=\underline{g}_3$ in which case α is arbitrary.) Hence the optimum indeed corresponds to a minimum of the error E. The minimum interpolation error is obtained from (1) - (3) as

$$E_m = (\underline{g}_2 - \underline{g}_1)^T (\underline{g}_2 - \underline{g}_1) - \alpha (\underline{g}_2 - \underline{g}_1)^T (\underline{g}_3 - \underline{g}_1) \; . \qquad (4)$$

Several comments are in order. The error E in (2) is equal to the square of the distance between the point $\underline{g}_2$ and any point $\underline{g}$ on the line given by (1) in the p-dimensional vector space. Figure 1 illustrates this for the case p=2. The point X on the interpolation line AC defines a vector $\underline{g}$, and produces an interpolation error equal to $(BX)^2$. Clearly, the distance (and hence the error) is minimum when the vector $(\underline{g}-\underline{g}_2)$ in (2) is orthogonal to the interpolation line (1). From (3), (4) and Fig. 1, one can show that $\alpha=AD/AC$, and $E_m=(BD)^2$. The minimum error E_m given by (3) is zero only when p=1 or $\underline{g}_2=\underline{g}_1$ or $\underline{g}_2=\underline{g}_3$. In all other cases, E_m is nonzero.

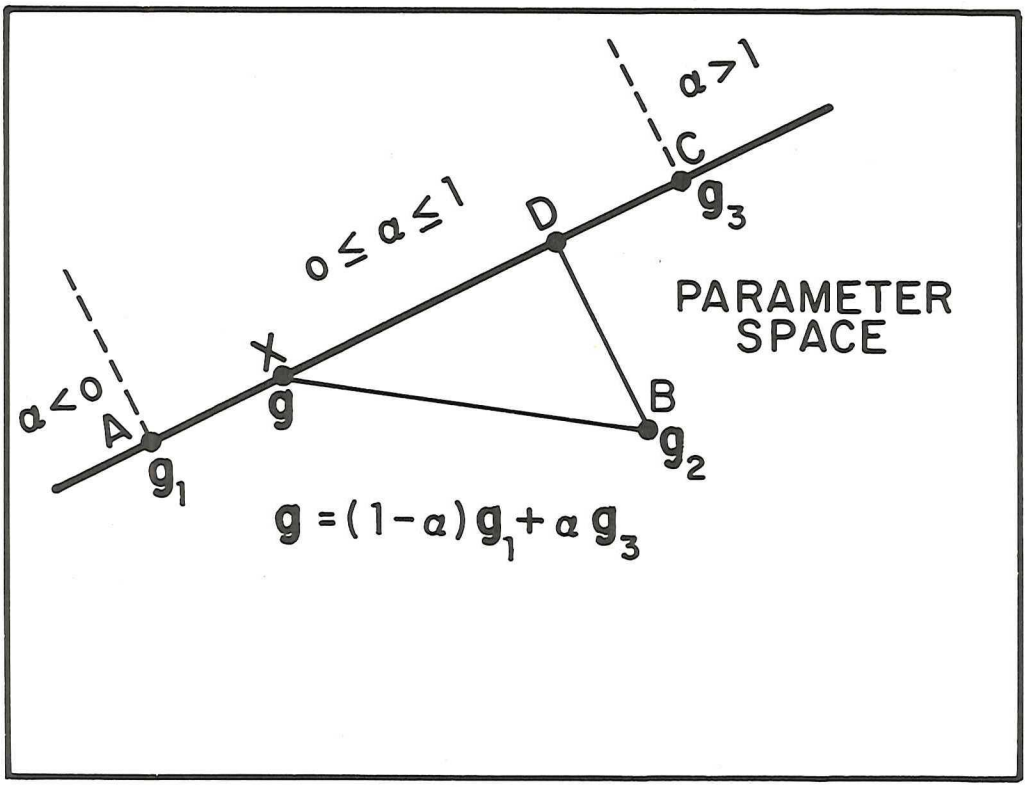

Fig. 1 Optimal linear interpolation in parameter space.

Another important result is that α and E_m do not depend on the relative frame locations in time. In fact, the above optimal interpolation problem can be stated in terms of g_1, g_2 and g_3 only, i.e., without considering time explicitly.

If we include time as an additional coordinate, then we have the (p+1)-dimensional space as shown in Fig. 2. Here, time is plotted along the X-axis and, for convenience, all the p parameters are lumped together and plotted along the Y-axis. The interval between successive data frames is assumed to be L seconds. The curve ABC represents the actual variation of parameters in time, while the line AC is used in the SLI scheme. The optimal interpolation point D is obtained by computing α from (3), determining the ordinate of the point on the line AC at time $(n+2\alpha)L$ seconds and marking D with this ordinate at time $(n+1)L$ seconds. The minimum interpolation error E_m in this case is given by $(BD)^2$. From this interpretation, we have the following: The SLI scheme corresponds to a line, while the OLI scheme produces a piecewise linear characteristic (ADC in Fig. 2). The latter will be closer to the actual parameter curve than the former. The piecewise linear variation in time has an interesting application to variable frame-rate transmission systems (see Section A.3). Also, if synthesizer parameters are to be updated at intervals less than L seconds, then this piecewise linear characteristic can be used to generate the additional interpolated parameters.

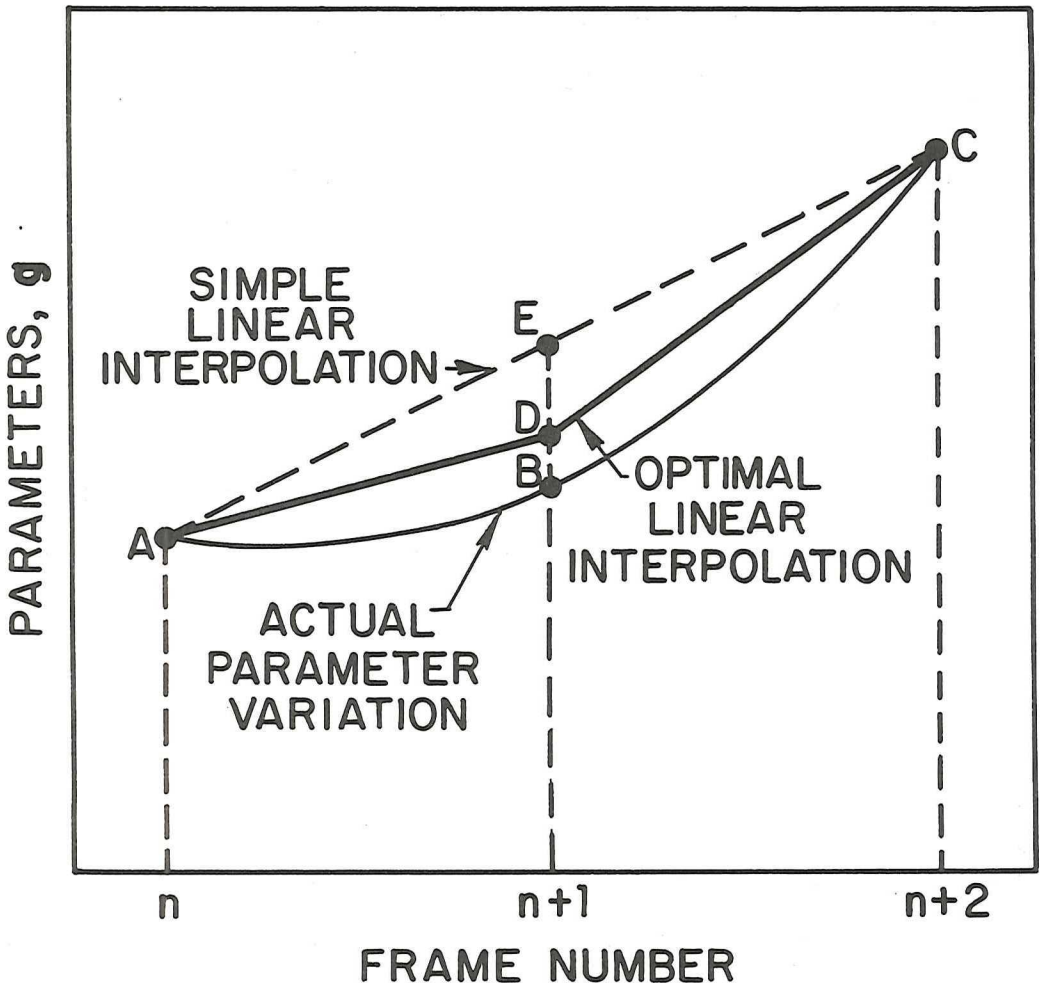

Fig. 2 Interpretation of optimal linear interpolation as a function of time.

As can be seen from (3), α can become negative or greater than 1. In order to limit the range of α for quantization purposes, when $\alpha>1$ we use $\alpha=1$, and when $\alpha<0$ we use $\alpha=0$. When $g_3=g_1$ we arbitrarily assign $\alpha=0.5$.

In the above development, we have implied that the vectors g_1, g_2 and g_3 represent the predictor parameters before quantization. However, the receiver has access only to the quantized parameters (i.e., after coding and decoding). In this case, α is computed from (3) after replacing g_1 and g_3 by their quantized values.

2. Choice of Parameters for Interpolation

The importance of a proper choice of interpolation parameters can be illustrated by the following example. Since we believe that accurate spectral representation of the speech signal is necessary for good quality synthesized speech, it seems on the surface that interpolation of points on the envelope of the log power spectrum is a reasonable thing to do. From a computational viewpoint, such an interpolation is expensive. But, let us leave that problem aside. For simplicity consider a spectrum with one formant (i.e., a pair of complex conjugate poles). Figure 3 depicts two such spectra (A and B) corresponding to two successive frames. The spectrum for the interpolated case should also have one formant that lies in between the peaks of the given two spectra. However, use of spectral interpolation results in a spectrum having two formants as shown by C in Fig. 3. The effect of such an interpolation on speech quality can clearly be disastrous.

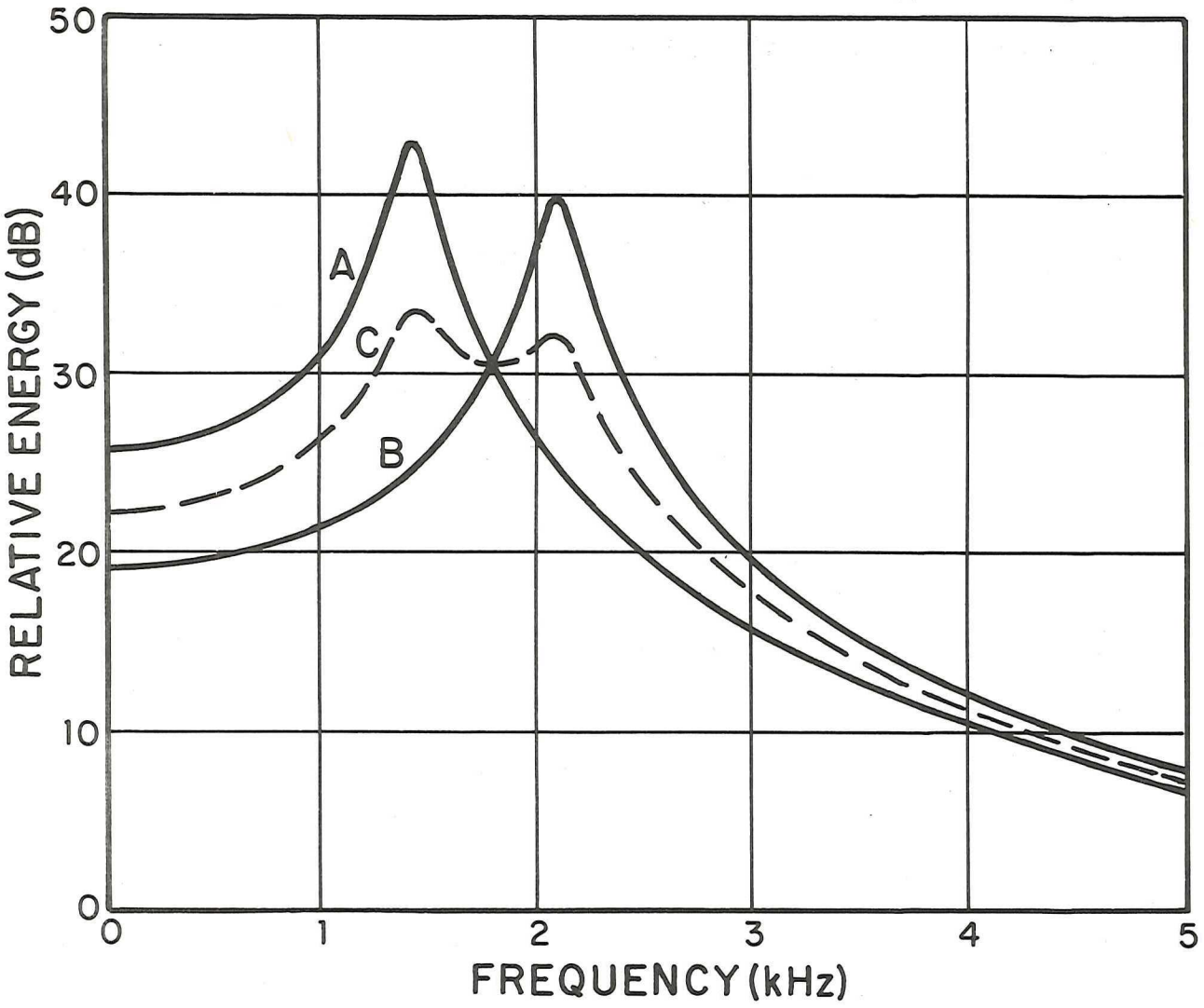

Fig. 3 Linear interpolation of log spectral values.

From the physics of the vocal tract (and also from the above example), it seems that the formant frequencies should be used for interpolation. This is also supported by the fact that formants, as seen on spectrograms, vary rather smoothly in continuous speech. For linear prediction analysis, poles of the predictor are most closely related to the formants. However, the problem of correctly identifying formants from poles is nontrivial. Also, all the poles of the predictor do not correspond to formants. It is not clear how to interpolate poles that are not formants. (Note that adjacent data frames can have different numbers of real and complex poles.) For these reasons, it is desirable to search for alternate sets of parameters for interpolation.

In our past investigations, we had used autocorrelation coefficients (first p+1 coefficients) of the speech signal, predictor coefficients, reflection coefficients and LARs for simple linear interpolation. Listening tests on the synthesized speech in these cases did not indicate any perceivable differences in quality. In view of the relatively flat sensitivity properties of LARs, we are using them as transmission parameters (BBN Report No. 2976). Hence, it is convenient to interpolate LARs directly. There is however another reason for choosing LARs for OLI. The error measure E given by (2) in this case is the LAR error i.e., total-squared error between the actual LARs and the interpolated LARs. We have shown in the context of quantization (BBN Report No. 2976) that the LAR error has a fairly linear relationship with the spectral error, where the latter is the sum over frequency of the absolute deviations in the log spectral values of the linear predictor. This

together with our assumption that an accurate representation of the envelope of the speech spectrum is necessary for good quality speech indicate that minimization of the LAR error due to interpolation should improve speech quality. This indeed has been our experience.

3. Application of the OLI Scheme

To explain some of the details in applying the OLI scheme, consider a fixed frame-rate system transmitting 50 data frames/sec, i.e., once every 20 msec. Assume that the receiver updates the synthesizer parameters every 10 msec (i.e., 100 times/sec). This requires one interpolation per data frame received, and hence transmission of one α per transmitted data frame. For computing these α's, the predictor parameters must be available once every 10 msec even though data will be transmitted only every 20 msec. From experiments, we found that 2 or 3 bits are adequate for quantizing α. From this example, this means that the transmission of α increases the total bit rate by 100-150 bits/sec.

In the example given above we assumed a time-synchronous (pitch-asynchronous) interpolation. For a pitch-synchronous OLI scheme, clearly analysis has to be done pitch-synchronously as also the transmission of α. As we have chosen to work with time-synchronous analysis and transmission (BBN Report No. 2976), we shall consider in this report only the time-synchronous OLI scheme.

Let us next consider the application of the OLI scheme to a variable frame-rate transmission system. For such a system, analysis is performed at a high rate, e.g., once every 10 msec.

However, parameters are transmitted only when the speech spectrum has changed sufficiently since the last transmission. The time interval between successive transmissions can vary, for example, from 10 msec to 80 msec. Before we discuss the different ways of applying the OLI scheme to the variable frame-rate system, let us define a couple of terms for the latter system. Analysis rate is, as the name suggests, the rate at which linear prediction analysis is performed. Basic (internal) rate is the rate at which the variable frame-rate transmission system looks at the analyzed data to decide when to transmit*. Denote the time intervals for these rates by T_a and T_b msec respectively. Let us illustrate the two definitions by considering an example where the two rates are not the same: Analysis is performed every 10 msec, while the extracted parameters are transmitted at variable intervals that are multiples of 20 msec i.e., analysis rate = 100 frames/sec and basic rate = 50 frames/sec.

A simple way of applying the OLI scheme is to compute and transmit α for every analysis frame not transmitted. α for a frame is computed as discussed in Section A.1 from the parameter vectors of that frame and of the two adjacent transmissions. For the example mentioned above, if the average transmission rate is 25 frames/sec then α will be transmitted at an average rate of 75 bits/sec (using 3 bits for quantizing α). Note that at least one α is transmitted between any two data transmissions. If, however,

---

*We tacitly assume that the (analysis-rate/basic-rate) ratio is an integer.

basic rate and analysis rate are equal for a system then whenever two transmissions are spaced T_a msec apart, no α will be transmitted.

If the criterion for detecting spectral changes in the variable frame-rate system is properly chosen, then only relatively small spectral changes occur between the data of a transmitted frame and the analysis frames which follow but lie before the next transmission. Hence it may be unnecessary to transmit α for every analysis frame not transmitted, as suggested above. Several alternative schemes can be suggested. We shall only consider the scheme where at most one α is transmitted between every two transmissions. Below we describe how α is computed.

We associate α with the frame that is more or less at the center between successive transmissions. If the number, say (N-1), of analysis frames between successive transmissions (say frames n and n+N) is odd, then α is associated with the analysis frame n+m where m=N/2. If (N-1) is even, then m can be arbitrarily taken as (N-1)/2. A simple way to compute α is to project the parameter vector for the frame (n+m) onto the line joining the parameter vectors for the frames n and n+N. However, in this scheme we are not making use of the data available for the frames n+i, $1 \leq i \leq N-1$, $i \neq m$. We can make use of these frames of data as follows. Select any α to interpolate for g_m, using the line between g_n and g_{n+N}. This defines a piecewise linear characteristic similar to the one shown in Fig. 2. From this piecewise linear characteristic, interpolated parameters for frames n+i, $1 \leq i \leq N-1$, $i \neq m$ can be

determined. Define the combined interpolation error as the sum of the total-squared errors between the actual and interpolated parameters for these (N-1) frames. The OLI problem then is to find α that minimizes this combined interpolation error. This minimization can be done analytically, but leads to an involved expression for α. We shall not give the details here, since this method did not produce any perceivable improvement in speech quality relative to the OLI scheme that uses only g_n, g_m, and g_{n+N} for computing α.

4. Experimental Results and Recommendations

We have incorporated the OLI scheme in our simulation of a linear predictive speech compression system. By synthesizing a number of speech utterances we investigated the differences in speech quality between using the OLI scheme and the SLI scheme. Both fixed and variable frame-rate transmission systems were used. To evaluate speech quality differences we used informal listening tests and carefully studied the waveform plots and spectrograms of synthesized speech, and the plots of spectral error due to interpolation. Often, differences (due to the use of one or the other interpolation scheme) observed in waveforms, spectrograms and spectral error plots suggested specific sections within the synthesized utterances for careful comparison in informal listening tests.

The results of our informal listening tests indicate that the OLI scheme improves speech quality during rapid transitions in the speech signal. When speech formants were changing rather smoothly and data transmission frame rate was sufficiently high (e.g. 50 frames/sec or more), the speech quality improvement, if any, was not apparent. When perceivable improvements in speech quality did occur, such as during rapid transitions and at low transmission data rates, listeners characterized them by one or more of the following terms: more clarity, less "muffled", absence of "pops" or "blips", and distinct or well-preserved transitions. It should be reiterated that the quality judgments were done relative to the case where the SLI scheme was used. With the latter scheme, either some low frequency noise or occasional "pops" and "blips" were produced in utterances containing a number of transitions (voiced-unvoiced) or stop sounds. When we used the OLI scheme, these effects either disappeared or were noticeably decreased. Both earphones and loudspeakers were used for listening. Some minute quality differences perceived when listening through earphones did not come through the loudspeakers, while some other effects were made less clear by loudspeakers. However, the low frequency noise and the "pops" and "blips" mentioned above were more pronounced when listening through loudspeakers.

Using the OLI scheme we found that a proportionately bigger quality improvement was obtained when the frame rate of transmission was lower. In another experiment, we investigated the effect of quantization accuracy of the interpolation parameter α on speech quality. We did not find any perceivable differences in speech

quality when we varied the number of bits used for quantizing α from 2 to 6 bits. Therefore, we suggest that 2 or 3 bits per interpolation coefficient should be adequate. This means that the use of the OLI scheme increases the total transmission rate by 50-150 bits/sec.

For the variable frame-rate system, we experimented with many ways of using the OLI scheme (see Section A.3). Differences in speech quality due to these different methods were mostly not perceivable. We recommend the OLI scheme that transmits at most one interpolation parameter per transmitted data frame, as it increases the total bit rate the least.

Waveform comparisons indicated that the OLI scheme produced a better timing of speech events (such as duration of stops and plosives) than the SLI scheme. An example is [b] in "little blanket". For this case, α at the start of [b] was close to 1 and at the end of [b] was close to 0, which resulted in a longer [b] than for the SLI scheme. The waveform of the natural speech also had a large duration for [b] in this example. Comparative studies using spectrograms showed that speech synthesized with the OLI scheme had its formants varying more smoothly in time than that with the SLI scheme. The variation of formants in the former case was found to be closer to that of the natural speech.

The spectral error between the interpolated parameters and the true or actual parameters is defined as the average over frequency of the absolute deviation in the log spectral values of the linear predictor when using the interpolated and the actual parameters. We

plotted spectral error as a function of time for both SLI and OLI schemes for a number of speech utterances. The major differences between the two plots occurred mainly during voiced-unvoiced transitions and rapid sonorant transitions. In these instances, the spectral error was much larger for the SLI scheme than for the OLI scheme.

In conclusion, we wish to make the following remarks. First, our past experience indicates that even though speech quality improvement due to any one factor such as improved interpolation, quantization, etc., may not be perceivable, the combined effect due to all of these differential improvements can be quite significant. This favors the use of the OLI scheme even in cases where it produces only a minimal improvement of speech quality. Our second remark deals with the question of computational burden at the transmitter resulting from the use of the OLI scheme. For a fixed frame-rate system, the OLI scheme requires computing the predictor parameters twice as many times if one interpolation coefficient per frame is transmitted. However, for a variable frame-rate system with equal basic and analysis rates, no extra predictor parameters need be computed. Thus, from a practical viewpoint, we recommend the use of the OLI scheme in a variable frame-rate speech compression system. For this system, we also recommend transmitting at most one interpolation coefficient per transmission, since this limits the increase in bit rate as well as amount of computation arising from the use of the OLI scheme.

B. Improved Pitch Quantization

Quantization of pitch presents an altogether different problem from the quantization of other transmission parameters. The major difference is that the decoded pitch values are constrained to be integers (samples per pitch period). Another difficulty arises in attempting to quantize the log pitch in that at the high frequency end (small pitch period) of the range of interest, the quantization bin size, as found by dividing the log pitch scale into equal segments, can be smaller than the distance between two allowable pitch values (for decoding). This leads to cases where two distinct quantization bins yield the same decoded value, thus wasting some quantization levels. In ARPA NSC Note #49, we proposed a method for deriving the pitch encoding and decoding tables in such a way that maximum usage is made of the different quantization levels. Our simulation system was modified to use this improved pitch quantization scheme. Statistics of differences in quantized pitch values using this scheme were collected for a number of speech utterances from male and female speakers for use in Huffman coding of pitch.

C. Real Time System

Our Dual Port SPS-41 was delivered on December 17, 1974. Except for the second analog to digital and digital to analog converters, we have now received all the equipment ordered from SPS. With the additional exception of the network interface, the hardware necessary for our NSC system has been delivered.

Acceptance tests supplied by SPS were run on the 41, and passed. However, ISI and SRI have discovered errors in the design of the 41 configuration used by the NSC project. We have confirmed the presence of these errors in our machine. SPS, Inc., is investigating these problems and when causes and solutions are determined, the modifications will be installed and checked in our machine before an attempt is made to modify the West Coast machines.

III. TABLE OF CONTENTS

Page

III. VOCODED-SPEECH QUALITY EVALUATION

1. INTRODUCTION

The ultimate criterion for determining the quality of the speech that is produced by any compression, encoding or transmission system is the way it sounds to the human listener. Given that several compression algorithms are being developed under ARPA auspices, each of which is likely to have several adjustable parameters, it is necessary to devise formal procedures for assessing the acceptability of the speech produced by different systems, or by different parameter selections for a given system.

Intelligibility is not likely to be a primary issue in such test procedures; we assume that all the systems that will be seriously considered for applied use will produce speech that is highly intelligible, at least in context and under typical conversational conditions. Speech quality, however, may differ considerably from system to system. Moreover, quality is not a simple unidimensional aspect of speech, so the speech from one system may be better than that from another with respect to some qualitative aspects but poorer with respect to others. It is important, therefore, not only to be able to rank order systems in terms of the overall quality of the speech produced but to devise tests that will provide some information concerning specific ways in which the quality of one sample of processed speech differs from that of another. The purpose for obtaining such information is not only to permit the comparing of the systems but also to provide data that can be used to improve the performance of any given system.

2. OBJECTIVES

The primary objectives of the work during the initial phase of this project were: (a) the selection of a set of speech materials (six sentences) that would provide an adequate basis for assessing how well candidate vocoding systems preserve the various features of speech that determine its quality, (b) the selection of several talkers whose speech represents desirable ranges of variation with respect to properties of interest, such as pitch, nasality, and word emission rate, (c) the specification of a set of candidate systems that could be used for preliminary evaluation exercises, (d) the preparation of the stimulus material to be used in listening tests, and (e) determination of the feasibility and sensitivity of alternative quality-assessment procedures.

3. SELECTION OF SPEECH MATERIAL

The objective in selecting speech material for use in evaluating vocoder performance is to choose material that will both contain reasonably complete representation of the individual speech sounds and also present these sounds in phonetic and prosodic contexts that will assure a stringent test of the vocoder's ability to preserve the naturalness of conversational speech. Before describing the specific material that was selected, it will be helpful to review the considerations that dictated the selection that was made.

3.1 Potential Sources of Quality Degradation

Inasmuch as the purpose of a vocoding system is to permit transmission of speech signals over a channel of severely limited bandwidth, there will be, by definition, less information (in the

negative entropy sense) in the output waveform than there was in the input waveform. The output will be a degraded version of the input. A successful vocoder system will retain, in the output waveform, just those aspects of the input signal that are most critical in speech. Any system that works in real-time will preserve the gross temporal pattern of the input. The other critical aspects of speech that must be preserved are (1) spectral information, (2) short-term temporal pattern, and (3) the fundamental frequency pattern.

There are several ways in which LPC vocoders can introduce discrepancies between the input waveform and the waveform as resynthesized by the vocoder. The vocoder operates by first selecting a short time-sample of the input waveform (e.g., 20 msec.), and matching the spectrum of the sample as accurately as possible with an all-pole approximation. The sample is viewed through a Hamming window to eliminate the sudden discontinuities at the start and end of the waveform sample. Errors can be introduced in two ways at this stage. First, the input spectrum may change markedly during the 20-msec. waveform sample, so that detail may be lost as a result of the averaging process. Second, the all-pole model may not be capable of adequately matching the input spectrum if the spectrum contains zeroes or more poles than the model can use.

After the spectral match has been obtained, the predictor coefficients that define the model spectrum are quantized. A third possibility for error is introduced by this quantization, and the larger the quantization step size, the greater the potential error.

The whole process is repeated with a new waveform sample, which is selected by advancing the window down the waveform by one frame. The window-advancing procedure is a fourth potential

source of error: If the frame size is larger than the width of the window, successive waveform samples will not overlap, and any acoustic events occurring between samples will not be analyzed.

Our choice of speech materials for testing LPC vocoders was strongly influenced by our desire to be able to associate identifiable deficiencies in the quality of vocoder outputs with the four possible error sources mentioned. Additional errors may be introduced by the pitch extractor during the measurement of fundamental frequency, and also by the quantization of the resulting pitch values.

The distinction among error classes is an important one conceptually because it suggests that any source of degradation may impose an upper limit on the quality of the speech produced. It would be pointless, for example, to attempt to reduce quantization error indefinitely if improvement in quality were being precluded by the inadequacy of the spectral match obtained from the model. And, conversely, if quantization error is sufficiently great, it may dominate irrespective of the adequacy of the spectral match. In this study we will not compare all-pole models with models that have both poles and zeroes. We will, however, be concerned with the effects on quality of the number of poles in the model, and of the quantization of the filter coefficients, and of the (effective) quantization of time by the window-size and frame rate.

These variables can undoubtedly be traded off against each other, within limits, without having large noticeable effects on intelligibility. That is to say, one might expect that for some range of these variables a "downward" change in one of them might be compensated by an "upward" change in the other, leaving the overall intelligibility of the speech about the same. However,

while such compensatory adjustments may leave the intelligibility of the speech unaffected, the resulting speech may not be equally acceptable to a listener. Suppose, for example, that a 14-pole approximation with relatively coarse quantization sounded like "computer speech," or the output of a tape recorder with speed variability, heard over a high-fidelity transmission system, whereas a 10-pole approximation with fine quantization sounded more like natural speech heard over a poor transmission system. It is likely that the perceived cause of the degradation could affect the acceptability of the speech to the listener. One may be more comfortable, for example, listening to what sounds like a person over a poor transmission system, than to what sounds like a computer over a good system.

3.2 Acoustic Properties of Speech

The speech sounds that occur in running speech fall into four broad categories: vowels and semivowels, nasals, fricatives, and stops and affricates.

a. Vowels and semivowels (/w,r,l,y/). These sounds are all (except /l/) made with the vocal tract relatively open, and the sound energy originates in the vibrations of the vocal cords; consequently, the sounds are relatively intense. They are further characterized by a spectrum that shows from two to five concentrations of energy (formants), that change relatively slowly with time (except during the release of /l/). The spectra of the vowels and semivowels (except /l/) can be accurately described by a set of resonances (poles) which effectively filter the energy injected at the bottom of the vocal tract by the vibration of the vocal cords. These sounds are characterized by a relatively simple and stable spectrum. The envelope of the signal changes only slowly.

b. Nasals (and /l/). The nasal sounds are also relatively intense. They are made by opening the velum, and closing the mouth cavity, during vibration of the vocal cords. Thus, in addition to the unconstricted passage through the nose, there is a closed side cavity which traps some of the energy, and results in the introduction of anti-resonances (zeroes) into the spectrum. The spectrum of /l/ contains zeroes also, due to the cavity formed behind the tongue tip. These sounds are characterized by a more complicated spectrum, and one that can change quite abruptly. However, as in the case of the vowels, the envelope of the signal changes only slowly.

c. Fricatives (/f, θ, s, š; v, ð, z, ž/) and /h/. Fricative sounds are made by forcing air through a narrow constriction. The constriction is above the vocal cords except for /h/, so there is a cavity behind the frication source, which introduces zeroes into the spectrum. Frication noise can occur with or without vibration of the vocal cords. The spectral characteristics of frication noise are very different from those of voiced sounds that are not fricatives. The build-up and decay of frication energy is usually gradual. The changeover between voicing excitation and frication excitation, at the junction of a vowel with a fricative, is one example of an acoustic event whose short-term spectral and temporal properties must be preserved for high intelligibility and quality. The envelope changes at these boundaries are also large.

d. Stops and affricates (/p, t, k, b, d, g, č, ǰ/). Stops and affricates involve a total closure of the vocal tract. During closure, air pressure builds up in the mouth, which causes a "pop" or burst when the closure is released. The burst, the subsequent aspiration of an unvoiced stop, and the frication of an affricate, all have spectral properties similar to those of fricatives. They differ from fricatives in that the duration of noise excitation is often extremely brief, and may include large and rapid changes

of the frequency of spectral peaks. Finally, after voicing begins there may also be rapid changes in formant frequency, lasting as little as 20-50 msec. Thus, stops provide examples of sounds whose spectra show sudden changes of both excitation type (noise vs. periodic) and of frequency, and large and rapid changes of envelope.

Speech materials for testing vocoder performance should include examples of as many as possible of these different types of acoustic events. Further, it is important to use sentence material rather than single words for at least part of the tests (but see Section 7.5 re phoneme-specific tests). The vocoding system will eventually be used to transmit conversations, which tend to occur in sentences, or at least phrases, rather than single words. Moreover, intelligibility and naturalness are strongly affected by prosodic features--pitch pattern and gross timing pattern including rhythm--which are much more salient in meaningful sentence material than in single words. Therefore, the sentences should exhibit a wide range of prosodic features, including several pitch contours and a range of patterns of stressed and unstressed syllables.

On the other hand, it is desirable to use as small a range of materials as possible, to keep the data-collection task within reasonable bounds. In addition, it would be helpful if the experimental results obtained permitted more than a simple rank ordering of the systems under test. If one could extract diagnostic information as well, the need for separate diagnostic testing could perhaps be eliminated.

3.3 Description of Selected Material

With the above considerations in mind, we composed a set of six short sentences. The first four sentences were intended to be diagnostic, and the last two were of a more general type. The

sentences were as follows:

1. Why were you away a year, Roy?
2. Nanny may know my meaning.
3. His vicious father has seizures.
4. Which tea-party did Baker go to?
5. The little blankets lay around on the floor.
6. The trouble with swimming is that you can drown.

3.3.1 General Comments

Prosody. To maximize the range of pitch contours represented, two of the six sentences were questions (#1 and #4). Emphatic stress was marked in two of the sentences (#2 and #4) by underlining a word. Words carrying emphatic stress are usually marked by a peak in the pitch contour, usually the major peak. Marking emphatic stress also clarified for the speakers the topic of the sentence, and tended to make all speakers read a particular sentence with the same (desired) intonation. Without this constraint, we obtained a reduced range of pitch contours, rather than an enlarged range, most sentences being produced with the same, neutral contour.

Rhythm. The sentences contain a wide range of patterns of stressed and unstressed syllables.

Phonetic balance. It is not clear that it is necessary, or even desirable, to include an occurrence of every English phoneme within our corpus, or to match the frequencies of occurrence to those that occur in the language as a whole. What is important is that every type of acoustic cue be represented, if possible in several different contexts. The latter aim was clearly met in our six sentences--and the former aim also was very nearly met,

although it was not one of our objectives. Despite the clearly nonrandom distribution of consonants in the first four sentences, the rank order by frequency of the consonants in our six sentences correlates highly significantly (Spearman's $\rho = 0.8$, $p < .01$). with the order for the language as a whole, as described by Denes (1963). All the English consonants appear in our six sentences, except one (/ʸj/). The balancing of the vowels is less good, the rank correlation obtained being nonsignificant. However, both in our six sentences and in the language as a whole, the two most common vowels (/ə, I/) appear three times as often as the third most frequent. The poor correlation for the vowels is due mainly to the over-occurrence in our sentences of some of the extreme vowels (/i, æ , a, u/) and diphthongs (/eI, aI, oU, aU/) at the expense of some of the intermediate vowels (/E, U, o/). We consider this departure desirable, since it tends to increase the amount of formant movements, thus providing a slightly more exacting test of vocoder performance than would be obtained with a corpus that better matched the distribution of vowels in the whole language.

3.3.2 Detailed Comments about Individual Sentences

Sentence 1: Why were you away a year, Roy? This sentence contains only vowels and semivowels (excluding /l/) and was spoken without a pause. Since none of the spectra associated with these phonemes contain zeroes, they should be approximated very accurately by an all-pole spectrum. In these cases, therefore, the basic assumption underlying LPC vocoders is valid, and the sentences give the opportunity for each vocoder to perform at its best. Furthermore, since levels and spectra change only slowly, the quality of the vocoded speech should be relatively insensitive to frame size. In other words, since Sentence 1 is an "all-pole sentence," the spectral approximation should be the best achievable with a given number of poles,

and any degradation of quality in the vocoded signal can be ascribed to losses due to quantization of the predictor coefficients.

Further, the pitch-detecting algorithms should also perform at its best. There are no very sudden changes in intensity, so pitch changes should be relatively slow and there are no voiced-voiceless boundaries, since all the sounds are voiced.

Sentence 2: Nanny may know my meaning. This sentence contains only nasals and nasalized vowels. As mentioned above, nasals have zeroes in their spectra, and nasalized vowels sometimes have a large number of formants, due to there being two separate resonance systems excited by the laryngeal pulses. Thus, Sentence 2 presents LPC vocoders with a class of sounds that should cause problems, mainly of the spectral-matching type. As in Sentence 1, all sounds are voiced, and changes in level are fairly gradual, so the pitch extraction algorithm should again perform at its best.

Sentence 3: His vicious father has seizures. This sentence contains vowels, and all the voiced and voiceless fricatives, except /θ/. Fricative spectra are very different from vowel spectra in that they tend to have a single broad energy concentration instead of three or four narrow ones. They also require zeroes in their spectra. The changes in spectra from vowel to fricative should be difficult for LPC vocoders to model, and the adequacy of the vocoded transition will also be influenced by the frame size. The transitions from unvoiced segments to voiced, and vice versa, should also test the ability of the pitch extractor to lock-on to the fundamental in difficult conditions. Voiced fricatives are notoriously problematical for pitch extractors, and their spectra may be difficult to match since they contain some aspects of both

voiced and voiceless sounds. No affricates were included, so that onset and offset of frication energy is gradual rather than abrupt.

Sentence 4: Which tea-party did Baker go to? This sentence contains vowels, and all the stops and affricates except /$\overset{v}{j}$/. These sounds are characterized by abrupt changes in intensity, and sudden changes between frication and voicing energy. The main acoustic cues to the identity of an initial stop consonant are the frequency of the burst, when the stop is released, and the formant transition into the following vowel. Both of these events are of short duration. To maximize the range of burst frequency and transition rates, a /t/ was placed both before a high front vowel ("tea") and before a low back vowel ("to").

This sentence should provide a critical test of a vocoder's frame size, since the spectra encountered are similar to those in Sentence 3. Thus, a system that performs adequately on Sentence 3 and badly on Sentence 4 can be diagnosed as having too large a frame size.

Sentence 5: The little blankets lay around on the floor; and Sentence 6: The trouble with swimming is that you can drown. The purpose of having two "general" sentences in the test, in addition to the preceding diagnostic sentences, was to include combinations of phonemes that have special acoustic correlates, and which would have spoiled the purity of the diagnostic sentences. Among the combinations deliberately included in Sentences 5 and 6 were a semivowel or liquid following an unvoiced initial consonant, since the first half of the second consonant tends to be unvoiced in this context (/fl/ in "floor," /tr/ in "trouble," /sw/ in "swimming," and /ty/ in "that you"). Examples of these consonants preceded by a voiced initial consonant were also included (/bl/ in "blankets,"

/dr/ in "drown"). Examples of syllabic /l/ can also fall in this class (/tl/ in "little," /bl/ in "trouble"). Second, combinations were included that should cause "formant splitting." This occurs when a vowel changes from being unnasalized to nasalized, as the velum opens in anticipation of a syllable-final nasal but before the mouth closes. Several of the formants of the vowel may split into two, as the new nasal resonances appear, and the vowel formants then disappear as the oral closure occurs. Thus, the number of formants may go from three or four up to six or more, and then back to two or three, which presents severe problems to an all-pole model. Formant splitting was not necessarily induced in Sentence 2, (Nanny may know my meaning), because every consonant was a nasal, and the whole sentence could be produced with the velum open. To insure that formant splitting occurs, it is necessary to precede the vowel-nasal combination with an obstruent. Several examples are included in Sentences 5 and 6 (/blæŋk/ in "blankets," /dan/ in "around on," /draUn/ in "drown," and perhaps /swIm/ in "swimming," and /kæn/ in "can," unless the vowel was so reduced that a syllabic /n/ resulted).

Third, some combinations of adjacent fricatives were included (/Θs/ in "with swimming," /zð/ in "is that"). A sequence of five rapid, unstressed syllables occurs in Sentence 6, which should provide a stringent test for even the best LPC vocoders.

4. RECORDING PROCEDURES AND TALKERS

Because the quality of the output of a speech-processing system can vary somewhat with the characteristics of the speaker's speech, recordings were made of 20 speakers (half of each sex), and six speakers were selected from the 20 so as to retain the desired range of speaker characteristics.

voiced and voiceless sounds. No affricates were included, so that onset and offset of frication energy is gradual rather than abrupt.

Sentence 4: Which tea-party did Baker go to? This sentence contains vowels, and all the stops and affricates except /j͡ʏ/. These sounds are characterized by abrupt changes in intensity, and sudden changes between frication and voicing energy. The main acoustic cues to the identity of an initial stop consonant are the frequency of the burst, when the stop is released, and the formant transition into the following vowel. Both of these events are of short duration. To maximize the range of burst frequency and transition rates, a /t/ was placed both before a high front vowel ("tea") and before a low back vowel ("to").

This sentence should provide a critical test of a vocoder's frame size, since the spectra encountered are similar to those in Sentence 3. Thus, a system that performs adequately on Sentence 3 and badly on Sentence 4 can be diagnosed as having too large a frame size.

Sentence 5: The little blankets lay around on the floor; and Sentence 6: The trouble with swimming is that you can drown. The purpose of having two "general" sentences in the test, in addition to the preceding diagnostic sentences, was to include combinations of phonemes that have special acoustic correlates, and which would have spoiled the purity of the diagnostic sentences. Among the combinations deliberately included in Sentences 5 and 6 were a semivowel or liquid following an unvoiced initial consonant, since the first half of the second consonant tends to be unvoiced in this context (/fl/ in "floor," /tr/ in "trouble," /sw/ in "swimming," and /ty/ in "that you"). Examples of these consonants preceded by a voiced initial consonant were also included (/bl/ in "blankets,"

/dr/ in "drown"). Examples of syllabic /l/ can also fall in this class (/tl/ in "little," /bl/ in "trouble"). Second, combinations were included that should cause "formant splitting." This occurs when a vowel changes from being unnasalized to nasalized, as the velum opens in anticipation of a syllable-final nasal but before the mouth closes. Several of the formants of the vowel may split into two, as the new nasal resonances appear, and the vowel formants then disappear as the oral closure occurs. Thus, the number of formants may go from three or four up to six or more, and then back to two or three, which presents severe problems to an all-pole model. Formant splitting was not necessarily induced in Sentence 2, (Nanny may know my meaning), because every consonant was a nasal, and the whole sentence could be produced with the velum open. To insure that formant splitting occurs, it is necessary to precede the vowel-nasal combination with an obstruent. Several examples are included in Sentences 5 and 6 (/blæŋk/ in "blankets," /dan/ in "around on," /draUn/ in "drown," and perhaps /swIm/ in "swimming," and /kæn/ in "can," unless the vowel was so reduced that a syllabic /n/ resulted).

Third, some combinations of adjacent fricatives were included (/Θs/ in "with swimming," /zð/ in "is that"). A sequence of five rapid, unstressed syllables occurs in Sentence 6, which should provide a stringent test for even the best LPC vocoders.

4. RECORDING PROCEDURES AND TALKERS

Because the quality of the output of a speech-processing system can vary somewhat with the characteristics of the speaker's speech, recordings were made of 20 speakers (half of each sex), and six speakers were selected from the 20 so as to retain the desired range of speaker characteristics.

The speech measurements that were made serve a secondary purpose in addition to that of speaker selection; they provide some data that will later be used to evaluate how well the speech-processing system extracts certain features from the speech signal, e.g., pitch, or the nasalization of nasal vowels.

4.1 Measurements Taken

Four aspects of the speech signal were recorded for each of the 20 individuals who comprised the set of candidate speakers.

4.1.1 Speech waveform: This was the standard analog representation of the sound pressure waveform, obtained from a boom-mounted electret microphone, as shown in Fig. 1 (bottom).

4.1.2 Glottal waveform: An analog signal was obtained from a miniature accelerometer (BBN 501) attached to the speaker's throat, as shown in Fig. 1 (top). The zero-crossing rate of this signal has been shown to be closely related to the fundamental frequency of the talker, and only slightly sensitive to the position of articulators in the higher portion of the speech-production apparatus (Stevens, Kalikow, & Willemain, 1974). It provides, therefore, the basis for an unambiguous determination of the fundamental frequency for voiced segments of speech. Pitch extraction is one of the subtler problems in vocoder performance; and the existence of a waveform produced in close proximity to the vocal folds, with a fixed time relationship to the speech waveform, can be used as a check on the performance of candidate pitch-detection algorithms operating on the speech waveform alone.

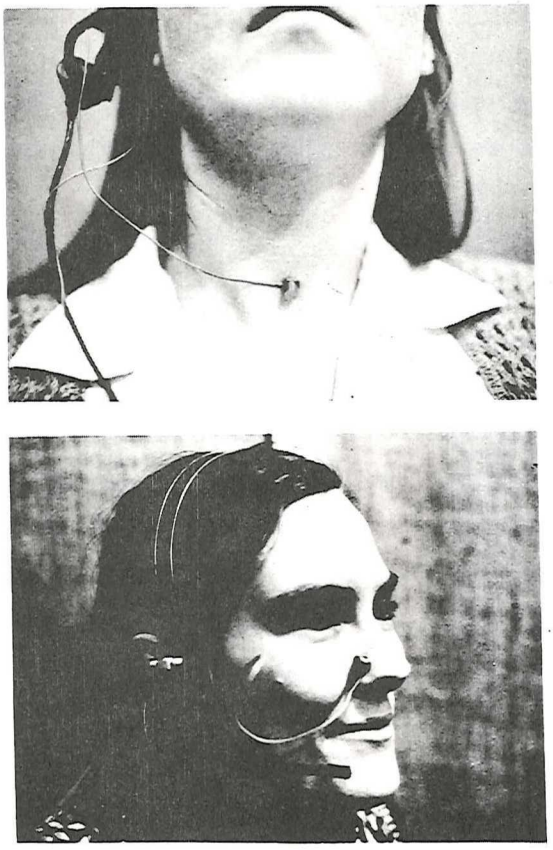

Fig. 1. Showing the accelerometer in position for detection of pitch and nasality (from Nickerson & Stevens, 1973).

4.1.3 A pulse-train representation of fundamental frequency: The analog signal described above was fed to a pitch-detection circuit that converted it into a pulse train, the leading edge of each 1-msec. pulse occurring at each positive-going zero crossing of the glottal waveform. This circuit has been used to produce a reliable indication of voice pitch in two BBN systems for on-line speech analysis and display (Kalikow & Swets, 1972; Nickerson & Stevens, 1973).

4.1.4 A measure of nasalization: An analog signal was obtained from a miniature accelerometer (BBN 501) mounted on the nose of the talker, as in Fig. 1 (bottom). This technique has been used to determine the degree of nasalization of specific speech segments produced by deaf and hearing speakers (Stevens, Kalikow, & Willemain, 1974; Stevens, Nickerson, Rollins, & Boothroyd, 1974). In brief, when attached to the nose, the accelerometer provides a signal that is an indication of the amount of acoustic coupling to the nasal cavity through the velopharyngeal port. The output, which is 10-15 dB higher when the velum is lowered--during nasalized sounds--than when it is raised, is fed to a component that rectifies and low-pass filters it and sends the result to the computer for further processing.

4.2 Recording Details

Two master tapes were recorded, one on a two half-track format machine (Braun TG1000), and the second on a four quarter-track format machine (Sony TC654-4). The Braun tape was used as a source for the vocoder evaluation test materials; the Sony tape provided a more detailed source for both speaker selection, and later vocoder optimization work. The voice signal was recorded directly on channel 1 of the Braun recorder. Its monitor output

was used in two ways: as input to the Sony's channel 1, and as a signal to drive a Ballantine 310A RMS voltmeter. The latter unit served as a level-maintenance meter for the talker. It was adjusted via the potentiometer to reach its center indication at the optimal recording level for speech peaks on the Braun record level meter. The Braun input control was adjusted for each talker such that speech peaks indicated 5 dB below 100 VU. Talkers attempted to maintain a consistent vocal effort throughout the session. Inexperienced talkers were monitored especially closely. The Braun frequency response is essentially flat throughout the speech range. Therefore, a high-quality signal was also recorded on the Sony, with the Braun merely serving as a microphone preamplifier.

Each of the twenty talkers recorded the sentences described above in a single session. Recording levels and signal quality were monitored and adjusted for each talker while the talker was familiarizing himself with the material by reading it aloud. The setting that most often needed adjustment was the gain for the nasalization channel, but all channels were monitored and adjusted so that peak levels rarely exceeded -3 dB re 100 VU on the recorder input meters. Once set during (nonrecorded) level-setting procedures, gains were not changed during the recording of a single subject.

Figure 2 shows the recording situation and the equipment used. The speaker was seated in a sound-shielded chamber. The voice signal was obtained from a head-mounted electret microphone (Thermo Electron Model 5336), which was enclosed in an open-cell foam windscreen, and placed at a fixed distance from the talker's lips and out of the breath stream. The two accelerometers were attached, one to the throat and one to the nose, with double-stick tape discs.

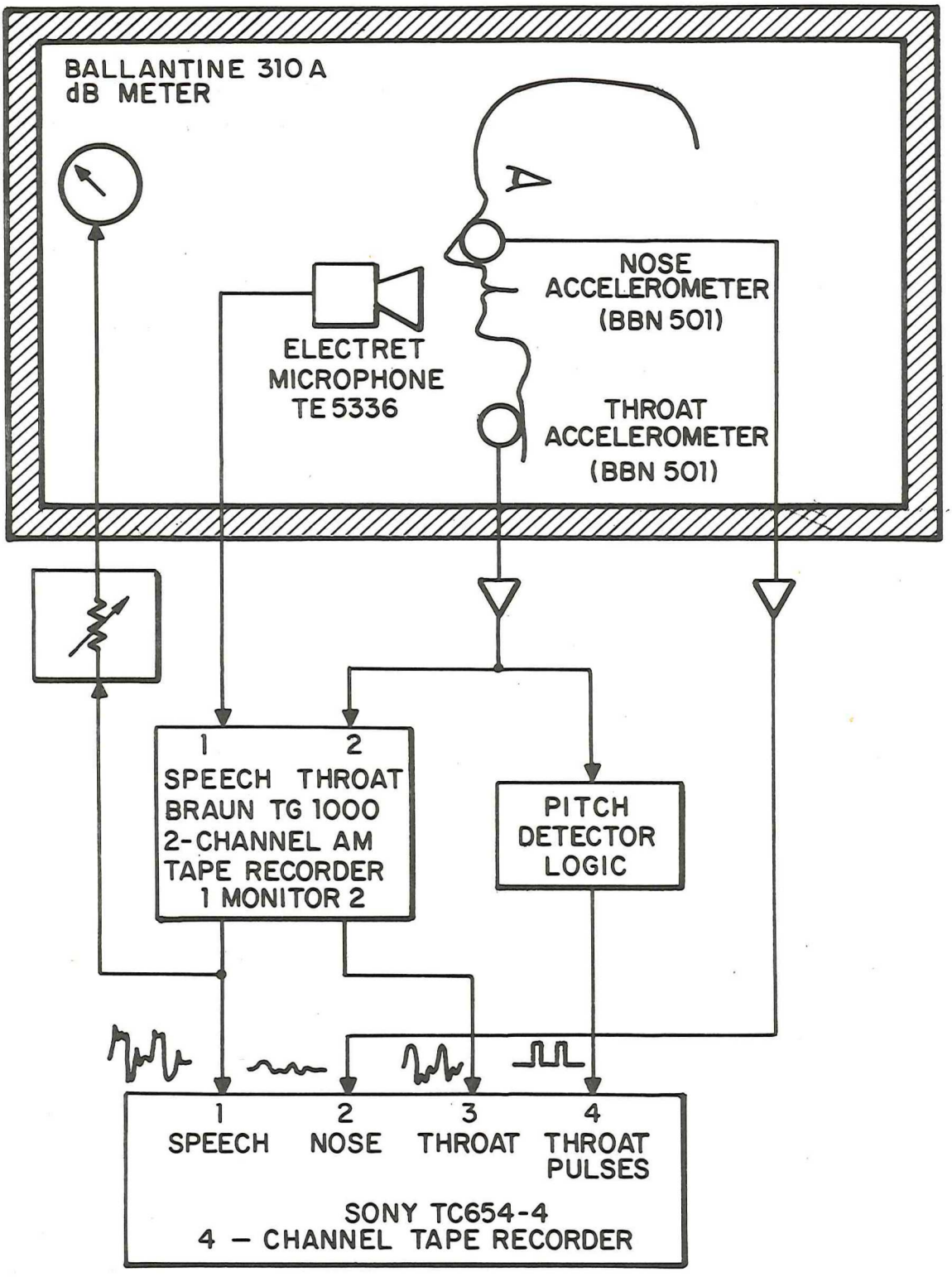

Fig. 2. Block diagram of recording arrangements.

The analog signal representing laryngeal excitation was obtained directly from a simple amplifier to which the throat accelerometer was connected. It was recorded on channel 2 of the Braun recorder, and the Sony channel 3 recorded the Braun monitor output. The analog laryngeal signal from the output of the first amplifier was also connected to the BBN pitch-detector logic, part of a larger interface for a PDP-8E computer. At the point in the circuitry where a pulse train is delivered to the output buffer, a connection was made to the Sony channel 4. The nasalization signal was amplified in the same way as the laryngeal signal, and was recorded on Sony channel 2.

4.3. Criteria for Talker Selection

The goal was the selection of three male and three female talkers whose speech covered the range of variation expected in potential users of the target vocoder systems. The strategy that was used was to select from our sample of twenty talkers three males and three females that collectively would represent a desirable range of values on each of three dimensions of interest: fundamental frequency, degree of nasalization, and speaking rate.

Prior to the taking of measurements, we carefully listened to the entire data tape. Some of the talkers were provisionally disqualified, due to speech mannerisms, accents, or other questionable signal qualities. While attempts had been made to limit the sample to talkers whose speech approximated "general American," certain speech sounds of a regional character were detected (especially #15, see Table 1 below). Furthermore, some talkers were noted to have unusually strident, or otherwise atypical, voices. These subjective data were kept for later use in the selection process. Objective data were collected on each of the twenty talkers for each of the three parameters of interest.

Table 1. Characteristics of the 20 speakers, with rank orders.

Speaker	Fundamental Frequency		Nasality		Duration seconds		Comments	Selected Speakers
Males	Hz	Rank	dB	Rank	Sec.	Rank		ID
1	148	10	16.6	7	2.80	9	slow, too inflected	
2	95	3	17.3	9	2.05	4=		DK
4	139	8	16.4	6	2.00	3		DD
5	97	4	18.5	10	2.20	6=		
6	118	6	14.3	1	2.20	6=	highly inflected	JB
7	143	9	15.2	3=	2.85	10	slow	
8	88	1	14.6	2	2.35	8	Boston accent	
10	119	7	16.0	5	1.95	1=	British accent	
17	101	5	16.7	8	2.05	4=		
20	91	2	15.2	3=	1.95	1=	Canadian accent	
Females								
3	167	2	15.2	2	2.00	1		AR
9	175	3	18.6	8	2.40	4=		
11	224	8	18.8	9	2.65	8	pauses	
12	212	7	17.7	7	2.70	9=	slow, pauses	
13	199	5	16.9	3	2.60	7	pauses	
14	185	4	17.0	4=	2.05	2		
15	246	10	23.1	10	2.45	6	Boston accent	
16	160	1	15.0	1	2.25	3	jerky	
18	232	9	17.5	6	2.70	9=	slow	PF
19	209	6	17.0	4=	2.40	4=		RS

4.3.1 Fundamental Frequency

The throat waveform track for each talker's utterance of Sentence 2 was played into the BBN speech analysis system. The average value of fundamental frequency for this utterance was calculated, using software developed for the production of hard copy of various speech parameters in another study (Nickerson, Kalikow, & Stevens, 1974). Additional parameter values of the distribution of F_0 points were computed, but the central concern here was the mean. The mean values of F_0 that were obtained are shown in Table 1. For the purposes of selection, the talkers are segregated by sex. Rank orders are presented with the data to give some idea of the relative spread and location of talkers. The six talkers that were finally selected are identified by an acronym that appears in the right-most column of the table, by which they will be referred to in what follows.

4.3.2 Nasalization

The analog signal recorded from the nasal accelerometer was input to a computer system that computes and displays a nasalization function over time. This function is produced with an arbitrary scale zero, and is displayed and averaged on a logarithmic scale. The scale zero is arbitrary because the amplitude envelope of the input is dependent not only on the accelerations of the talker's nose, but on the channel gain and transducer placement. Consequently, the most valid measurements of nasalization are those made within talkers rather than across talkers. Within-talker indices of nasalization strength have been defined for the purpose of assessing the degree of nasalization of defective speech (Stevens, Nickerson, Boothroyd, & Rollins, 1974). These indices are also appropriate for assessing the degree of nasality of normal speech, however,

and one of them was used here. This measure compares the degree of nasalization of sounds that are supposed to be nasalized with that of those that are not. The differences are usually in the range of 15 to 20 dB. A smaller difference indicates either a hyper- or hyponasal voice quality. The nasality of a speaker was defined for present purposes as the difference in the nasalization function measured in Sentences 2 and 4.

Each talker's tape was input to the system with gain settings that were constant for the two utterances. The first utterance was Sentence 2, "Nanny may know my meaning." This typically produces an elevated nasalization function exhibiting peaks at the syllabic nuclei. The maximal value of the function was recorded, irrespective of where it occurred within this utterance. The second utterance was Sentence 4, "Which tea-party did Baker go to?" This has no nasal phonemes; consequently, the nasalization function should be consistently low for a given talker. A "typical nonnasalized phoneme" value for this function was obtained by measuring, for each syllable in Sentence 4 what the largest value of the nasalization function was, within each of the nine syllables, and averaging them. The nasality criterion value for each talker was taken as the difference between this average and the nasal peak of Sentence 2, expressed in dB.

Note that this measure does not relate to the absolute level of nasalization to be expected in any talker's speech. It reflects rather the relative nasalization level within a given talker. The fourth column of Table 1 gives the nasalization value for each talker.

4.3.3 Speaking Rate

The time taken by each talker to produce Sentences 5 and 6 was measured with a stopwatch with the tape playing at half speed. Since the utterances were the same for each talker, there was no need to compute rates per syllable, and the average duration was used as an index of speaking rate.

The sixth column of Table 1 gives the duration data for each talker. Speaking rate was the only parameter of the three where the range was reduced rather than maintained in our sample. We excluded most of those speakers whose speech was slow, for two reasons. First, slower speech is never harder to code than fast speech, and, second, conversational speech tends to be faster than recitation form rather than slower.

4.4 The Talkers

The selection of talkers involved some compromising. More importance was attached to the desirability of covering the F_0 range, with the less important parameters being nasalization and speaking rate, in that order. This priority ordering is apparent in the choices shown in Table 1.

5. SELECTION OF CANDIDATE SYSTEMS AND PROCESSING OF SPEECH SAMPLES

The "systems" that were selected for initial testing were defined by twelve different combinations of parameter values for the BBN LPC system. In defining these systems, the bit rate was held constant at roughly 2600 BPS, while each of the following parameters was varied: frame size, quantization step size, number of poles in the model and nature of the transmission schedule (fixed or variable rate).

It was not necessary to do a factorial experiment with the several values of interest on each parameter. The purpose of the initial experiments was not to identify the best of the systems being compared, but, rather, to attempt to develop effective evaluation procedures. What was desired of the set of candidate systems was that it represent a reasonable range of quality, but that some of the systems be sufficiently similar in quality to assure a difficult judgmental task.

Table 2 lists the systems (combinations of parameter values) that were selected for initial experimentation. In addition to the speech produced by the twelve experimental systems, two types of control-condition stimuli were used: unprocessed speech and speech that was processed but not quantized.

Four hundred and sixty-eight sentences (6 sentences x 6 talkers x 13 systems--including one control) were processed. The processed sentences were then recorded on a master tape, together with an unprocessed version of each sentence.

6. EXPERIMENTATION

Two formal listening experiments have been done. A third has been designed and is about to be conducted. The completed experiments and their results are described briefly here.

6.1 Experiment 1: Triplet Comparisons

Task. The listener's task on each trial in this experiment was to decide which of three processed speech samples sounded most similar to a nonprocessed sample, and which sounded least similar. All four samples heard on a given trial were of the

Table 2: List of Parameters of Systems for Evaluation.

System No	No of Poles	Frame Size (msec)	Var. Rate Threshold dB.	Quantization Step Size dB.	Bit - Rate Expected	Bit - Rate Obtained
1	12	20		1.0	2650	2630
2	10	20		0.6	2650	2633
3	14	20		1.4	2700	2681
4	12	25		0.45	2640	2610
5	14	25		0.7	2640	2612
6	10	25		0.2	2680	2652
7	10	15		1.75	2666	2618
8	12	10	1.5	0.5	2660	2574
9	12	10	1.0	1.0	2650	2652
10	12	10	1.75	0.25	2627	2687
11	14	10	1.5	0.6	2685	2766
12	12	15	1.5	0.4	2600	2535
13	14	10	--- VOCODED BUT UNQUANTIZED ---			
14	---------- ORIGINAL WAVEFORM, DIGITIZED AND RECONSTITUTED					

same sentence, produced by the same talker. Only the systems by which they were processed differed.

Method. The trial procedure was as follows: A nonprocessed sentence was presented, followed by the three processed versions of the same sentence. Listeners were instructed merely to listen to the four sentences on the first presentation. The sentences were then heard in the same order and this time the listeners were required to indicate by pressing appropriate buttons which of the three processed sentences was most similar and which was least similar to the unprocessed sentence. The unprocessed sentence always was the first of the four sentences heard on a given presentation. Lighted response buttons marked the time of presentation of the standard speech and of each of the three processed versions. The stimulus triplets were recorded on tape cassettes which were operated under the control of a PDP-8 computer.

Nineteen college students in five groups served as test listeners. Each group had four listeners normally, although about two thirds of the time fewer than four showed up for scheduled sessions. Data were collected in half-hour sessions, in which from two to four listeners were tested simultaneously in an anechoic chamber. Each session consisted of 72 trials. The computer tallied the votes from all listeners on each trial, weighting "most similar" judgments plus one, "least similar" judgments minus one, and the remaining speech sample, zero. At the end of the evaluation session a summary of the judgments obtained during the session was printed out automatically by the computer. Figure 3 shows an example of a session summary produced by the computer immediately following the session.

```
SESSION: 10

# OF OBSERVERS: 3
OBSERVER     1 # 17
OBSERVER     2 # 19
OBSERVER     3 # 20

VOCODER SYSTEM A #:1
VOCODER SYSTEM B #:10
VOCODER SYSTEM C #:12
READY?
READY?
```

OVERALL SUMMARY OF JUDGMENTS

SYSTEM	# FIRST	# SECOND	# THIRD	SCALE VALUE
1	75	68	65	10
10	46	74	89	-43
12	88	64	59	29

JUDGMENTS FOR OBSERVER 17

SYSTEM	# FIRST	# SECOND	# THIRD	SCALE VALUE
1	24	24	21	3
10	16	25	29	-13
12	29	19	22	7

OF CONSISTENT JUDGMENTS: 70

JUDGMENTS FOR OBSERVER 19

SYSTEM	# FIRST	# SECOND	# THIRD	SCALE VALUE
1	26	22	24	2
10	16	25	30	-14
12	29	24	18	11

OF CONSISTENT JUDGMENTS: 52

JUDGMENTS FOR OBSERVER 20

SYSTEM	# FIRST	# SECOND	# THIRD	SCALE VALUE
1	25	22	20	5
10	14	24	30	-16
12	30	21	19	11

OF CONSISTENT JUDGMENTS: 51

Fig. 3. Example of printout of triplet data.

A different triad of vocoder systems, chosen from those shown in Table 2, was evaluated in each data-collection session for each group in counterbalanced order. The major body of data consisted of six sessions per group, of which four sessions were used to evaluate four independent triads from the twelve systems, and the other two were used to evaluate dependent triads. Additional data (about 6 more sessions) were collected with nonstandard triads, and as replications of triads.

An incentive system in which subjects were rewarded for consistent responses kept the interest reasonably high. Mechanical failures were rare. Tape synchronization and electronic switch failures were noted on about a dozen occasions of the approximately 20,000 tape plays.

Results. Only 12 of the 19 listeners participated in the 6 sessions necessary to hear all 6 system triplets. The data that were obtained from the remaining 7 listeners who heard some subset of the triplets are not presented here.

Inasmuch as each triplet occurred twice during a listening session, it was possible to compare both systems and listeners with respect to the degree to which the quality judgments that were made were consistent across the two occurrences. Table 3 shows the number of consistent judgments for each listener-system combination. Given the method that was used to calculate consistency, 108 represents the maximum possible score, and 36 represents chance. It is clear from the numbers in the table that consistency was not, in general, great. This bears out the impression of the experimenters that the perceptual task was a very difficult one; most of the systems that were being compared did not differ greatly in terms of quality.

Table 3. Number of judgments that were consistent between replications 1 and 2, by subject and by triplet. Maximum possible value for one cell = 108. Chance value = 36. The ringed data also appear on Fig. 3.

Subj.	3-9-10	1-2-4	1-10-12	1-9-11	5-6-12	7-8-11	Σ	Mean	Rank
10	68	57	70	70	50	72	387	64.5	1
17	73	65	(70)	72	52	42	374	62.3	2
2	63	50	74	69	67	50	373	62.2	3
3	68	65	64	68	55	51	371	61.8	4
19	59	57	(52)	56	56	45	325	54.2	5
20	72	52	(51)	41	55	45	316	52.7	6
5	56	47	52	47	51	27	280	46.7	7
1	45	47	48	52	37	33	262	43.7	8
11	47	50	36	21	21	25	200	33.3	9
7	32	30	22	26	40	39	189	31.5	10
6	37	27	31	27	35	18	175	29.2	11
9	35	39	15	6	23	37	155	25.8	12
N=12	655	586	585	555	542	484			
$\bar{x}$	54.58	48.83	48.75	46.25	45.16	40.33			
Rank	1	2	3	4	5	6			

Table 4 shows how each listener rank ordered the systems in each triplet when the results of the 72 trials of a given listening session were pooled. A plus sign indicates that the speech produced by a system was judged to be most similar to the unprocessed standard, a minus sign indicates that it was judged to be least similar, and zero represents an intermediate degree of similarity. The numbers at the bottoms of the columns indicate the difference between the number of plusses and minuses in the column. The closer this number is to 12, the greater the degree of consensus that this system produced speech that was most (if the number is positive) or least (if negative) similar to the standard.

The fact that most of these numbers are relatively small suggests again that the judgments were very difficult. There was complete agreement across the subjects on only two systems, within the triplets (5, 6, 12) and (7, 8, 11). System 6 was invariably chosen as worst within the context of the first triplet, and System 7 was always worst in the context of the second triplet. (Note, however, that 5 and 12 were relatively indistinguishable, as were 8 and 11.) Inspection of Table 2 suggests that the fact that both 6 and 7 had only 10 poles, whereas the systems with which they were compared had at least 12, may have been an important factor in determining their poor showings. The only other 10-pole system that was included in the sample (#2) also did rather poorly; it was judged as the poorest system within the triplet (1, 2, 4) by 8 of the 12 listeners, and as best by only 2. Again, the systems with which it was compared had at least 12 poles.

The data obtained in this experiment are being subjected to further analyses in an attempt to determine whether there were any sentence-specific or talker-specific effects. The first-order

Table 4. Pooled rankings of each triplet by each of 12 subjects.
The ringed values represent the data appearing in Fig. 3.

Listener	(3 9 10)			(1 2 4)			(1 10 12)			(1 9 11)			(5 6 12)			(7 8 11)		
10	−	0	+	0	−	+	+	0	−	0	−	+	0	−	+	−	+	0
17	−	+	0	0	−	+	0	−	+	−	0	+	+	−	0	−	+	0
2	0	+	−	+	0	−	0	−	+	0	−	+	0	−	+	−	+	0
3	0	−	+	−	0	+	+	0	−	+	0	−	0	−	+	−	+	0
19	−	+	0	+	−	0	0	−	+	−	0	+	+	−	0	−	0	+
20	0	+	−	+	−	0	0	−	+	−	0	+	0	−	+	−	+	0
5	+	0	−	+	−	0	0/−	0/−	+	0	−	+	+	−	0	−	0	+
1	−	+	0	0	+	−	−	0	+	+	0	−	0	−	+	−	0	+
11	−	+	0	+	−	0	0	+	−	−	0	+	0	−	+	−	0	+
7	+	−	0	+	−	0	+	0	−	−	0	+	+	−	0	−	0	+
6	0	+	−	+	−	0	0	+	−	0	−	+	0	−	+	−	0	+
9	+	−	0	−	+	0	0	+	−	0	−	+	+	−	0	−	0	+
	-2	4	-2	5	-6	1	1.5	-1.5	0	-3	-5	8	5	-12	7	-12	5	7

conclusion that we have come to, however, as a result of running the experiment is that this method is probably not an effective one to apply--atleast not without some modification--unless the qualitative differences among the systems being compared are relatively large. The advantage of the method is the fact that it is automated and it permits rapid data acquisition. It appears not to be sufficiently sensitive, however, to discriminate small differences in speech quality. It is possible that more consistent judgments might have been obtained, had the listeners been permitted more time to listen and relisten to the stimuli.

A second limitation with the triplet-comparison method is the fact that the procedure permits a rank ordering of only three systems at one time. If one's objective is to compare the quality of three specific systems, on an ordinal scale, then this is not a limitation. If one wants to rank order a larger set of systems, however, the procedure is probably not an efficient one to use. To rank order a larger set of items using the triplet-comparison procedure, one could conduct sessions with overlapping sets of triplets and then, perhaps, infer rank order within the union of the test sets. Given systems 1, 2, 3, 4, 5, and 6, for example, one might conduct listening tests with triplets (1, 2, 3), (4, 5, 6), (1, 3, 5) and (2, 4, 6). If the individual tests yielded the rank orderings (1, 3, 2), (6, 4, 5), (1, 3, 5), and (2, 6, 4), then one could infer that the rank ordering of the items in the union is 1, 3, 2, 6, 4, 5. If, however, the third triplet had yielded (1, 5, 3), there would be no way to order the union so as to be consistent with the individual sets, because two mutually exclusive orderings would be implied, one in which 3 is preferred to 5, and the other in which 5 is preferred to 3. Of course, to the extent that the judgments reflect the positions of the items on an absolute quality scale, such intransitivities should not occur.

6.2 Experiment 2: Rank Ordering

The procedure followed in Experiment 2 was designed to overcome some of the limitations that appeared to characterize the method used in Experiment 1. It was similar in some respects to a listening procedure that had been used to judge certain qualitative aspects (e.g., degree of nasalization) of the speech of the deaf (Stevens, Nickerson, Rollins, & Boothroyd, 1974).

Task. The listener's task on each trial in this experiment was to rank order the outputs from the 12 systems (plus one unprocessed and one processed but unquantized sample) after listening to the samples as often as he wished.

Method. The 504 sentences were dubbed from the master tape onto Language Master cards. (The Language Master is a tape recorder that uses short lengths of 1/4-inch magnetic tape, mounted on cards, as its recording medium.) The Language Master was first modified to reduce the level of AC hum, and to permit the signal to be recorded directly, bypassing the microphone.

An attenuator between the tape recorder and the Language Master was adjusted so that every sentence had the same maximum level, as measured by a HP427A AC voltmeter. Great care was taken to choose a recording level that both maximized the signal-to-noise ratio of the recorded signal, and also minimized clipping of the waveform peaks. This caution was necessary because preliminary test recordings showed that the "peakiness" of the vocoded signals was sometimes as much as 10-12 dB more than the unprocessed signal, as a result of its being minimum phase (Makhoul & Wolf, 1972, p. 93).

Each sentence started at the same point on its card, to minimize extraneous variability between stimuli. A photoelectric device was built for ensuring that each sentence began at the same point. The master tape was positioned so that the sentence to be dubbed began a fixed distance before the playback head. A card was placed in the Language Master, and as it started to more, it interrupted a light beam, which in turn operated the remote control of the tape recorder.

Each card was marked at its upper left-hand corner with a code that identified the speaker and sentence number, and a randomly assigned identification number. There were no clues on the card about which particular vocoder had processed the sentence: this information could be obtained only with the aid of a code table, not accessible to the subjects.

6.3 Order of Presentation

After the cards had been recorded and checked, the 14 cards for each sentence by each speaker were ordered by their (random) identification numbers, and held together with a rubber band. Ideally, the order in which a subject encountered the sentences should have been derived from a Graeco-Latin square that is also a double Williams square (Williams, 1950). This would have fully counterbalanced all possible serial position effects and sequential effects between pairs of sentences and pairs of speakers. Such sequence effects have been shown to be important between adjacent stimuli (Huggins, 1968): their occurrence between blocks of stimuli was considered likely enough to warrant a design that eliminates them.

Unfortunately, no Graeco-Latin square of order 6 exists. Therefore, three pairs of squares were constructed, which provide a complete counterbalancing of sequence of occurrence (each talker precedes each other talker an equal number of times) for both talkers and sentences, a complete counterbalancing of position of occurrence (early vs. late) for talkers, and an almost complete counterbalancing of position for sentences. The design is summarized in Table 5.

Each listener was given a copy of one matrix from Table 5 which specified the sequence of talkers and sentences to be followed. He first played through the specified pack of 14 cards, in order of ascending (or for alternate pairs of subjects, descenting) identification numbers and placed the cards in a "toast rack" in an approximate order from best to worst. The listener then listened to each of the sentences, as often as he wanted, rearranging their order until he was satisfied that he had the order that was indeed from best to worst.

Results. We have just begun collecting data with this procedure. Three listeners have been run, and the data from two of them have been partially analyzed.

The data that resulted from a single subject were 36 rank orderings of the 14 systems, each ordering representing a unique combination of one of the 6 sentences and one of the 6 talkers. These data were summarized in several ways: mean ranks were computed over sentences for each talker-system combination, over speakers for each sentence-system combination, and over sentences and talkers for each system. Figure 4 shows the means over sentences and speakers for each system; the results for one listener are plotted against those for the other.

Table 5. Counterbalanced design for order of presentation of speaker-sentence combinations, for rank-ordering task. Each subject followed through one of the six matrices.

	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
1	AR1	DK6	JB1	RS2	PF3	DD5	RS5	PF6	DD2	AR4	DK3	JB4
2	JB6	PF5	RS6	AR4	DD2	DK1	AR1	DD5	DK4	JB3	PF2	RS3
3	DK2	JB4	PF2	DD6	RS1	AR3	DD3	RS4	AR6	DK5	JB1	PF5
4	RS5	DD1	AR5	JB3	DK4	PF6	JB6	DK1	PF3	RS2	DD4	AR2
5	DD3	AR2	DK3	PF1	JB5	RS4	PF4	JB2	RS1	DD6	AR5	DK6
6	PF4	RS3	DD4	DK5	AR6	JB2	DK2	AR3	JB5	PF1	RS6	DD1
1	DK1	JB2	RS3	PF4	DD6	AR2	PF1	DD3	AR5	DK4	JB5	RS6
2	PF6	RS1	AR5	DD3	DK2	JB1	DD6	DK5	JB4	PF3	RS4	AR2
3	JB5	PF3	DD1	RS2	AR4	DK3	RS5	AR1	DK6	JB2	PF6	DD4
4	DD2	AR6	JB4	DK5	PF1	RS6	DK2	PF4	RS3	DD5	AR3	JB1
5	AR3	DK4	PF2	JB6	RS5	DD4	JB3	RS2	DD1	AR6	DK1	PF5
6	RS4	DD5	DK6	AR1	JB3	PF5	AR4	JB6	PF2	RS1	DD2	DK3
1	JB3	RS4	PF5	DD1	AR3	DK2	DD4	AR6	DK5	JB6	RS1	PF2
2	RS2	AR6	DD4	DK3	JB2	PF1	DK6	JB5	PF4	RS5	AR3	DD1
3	PF4	DD2	RS3	AR5	DK4	JB6	AR2	DK1	JB3	PF1	DD5	RS6
4	AR1	JB5	DK6	PF2	RS1	DD3	PF5	RS4	DD6	AR4	JB2	DK3
5	DK5	PF3	JB1	RS6	DD5	AR4	RS3	DD2	AR1	DK2	PF6	JB4
6	DD6	DK1	AR2	JB4	PF6	RS5	JB1	PF3	RS2	DD3	DK4	AR5

Several observations are worth making about Fig. 4. First, both listeners invariably gave the unprocessed speech sample the highest rating; all of the processed speech samples, including those that were not quantized, were readily distinguished from the natural speech. Second, the listeners were in quite good agreement in their rank orderings; the rank-order correlation (Spearman rho) was 0.8 ($p < .01$) in spite of the fact that about half of the scores were tightly clustered in one region of the scale. Third, excluding the unprocessed and the unquantized samples, the systems seem to fall readily into three quality classes: [1, 4, 5, 12], [2, 3, 6, 8, 9, 10, 11], and [7]. Table 6 shows this partitioning in terms of the parameters of these systems.

The thing that one immediately notices about this clustering is the fact that the best systems differ from the worst one in terms of both number of poles and quantization step size. It is also the case that most of the systems in the middle quality category have <u>either</u> fewer poles <u>or</u> larger quantization step size than those in the best category, or they have a variable transmission rate. These results are highly tentative, of course, but they do provide suggestions for future listening tests.

What is encouraging about these results is the good agreement between the two listeners whose data have been analyzed. It obviously would be imprudent, however, to decide that the procedure has great utility until we collect data from several more listeners and see whether the present high degree of consistency is maintained.

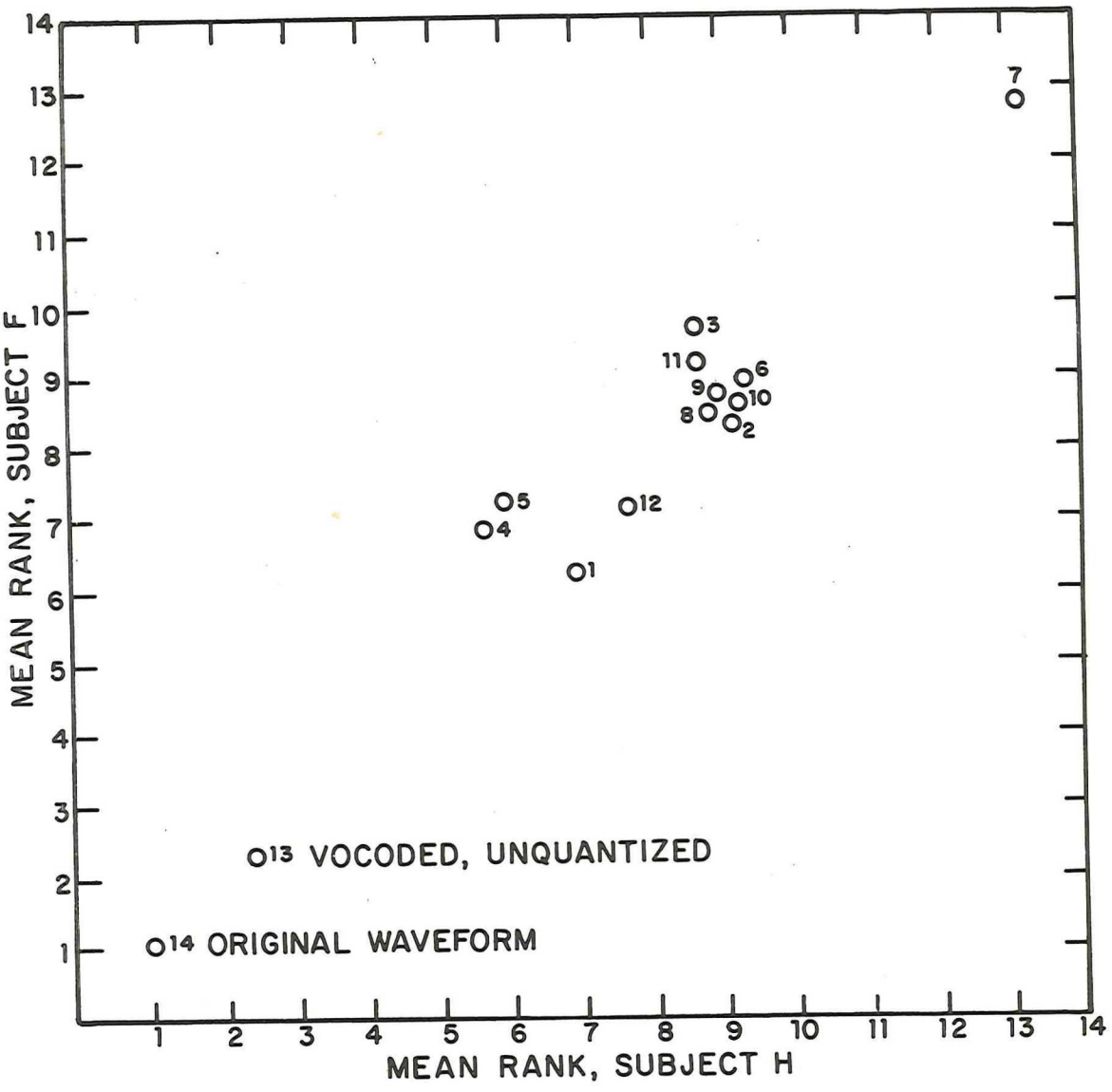

Fig. 4. Comparison of mean rank orders of fourteen systems by two subjects.

Table 6. Clustering of systems suggested by Fig. 4.

	System	Frame Size (msec.)	Quant. Step (dB)	Poles	Variable Rate Threshold (dB)	Bits/Sec.
Best	1	20	1.0	12		2630
	4	25	0.45	12		2610
	5	25	0.7	14		2612
	12	15	0.4	14	1.5	2535
Medium	2	20	0.6	10		2633
	3	20	1.4	14		2681
	6	25	0.2	10		2652
	8	10	0.5	12	1.5	2574
	9	10	1.0	12	1.0	2652
	10	10	0.25	12	1.75	2687
	11	10	0.6	14	1.5	2766
Worst	7	15	1.75	10		2618

7. FUTURE PLANS

Plans for the future include the following activities.

7.1 Collection of More Listener Data Using the Rank-Order Procedure

We plan to collect data from at least six subjects (four in addition to the two we already have).

7.2 Analysis of the Existing Data for Sentence-Specific and Talker-Specific Effects

The sentence material and the talkers were chosen to represent a broad spectrum of speech sounds and voice characteristics, because of the possibility that vocoders producing speech of similar overall quality might differ with respect to their ability to cope with specific sounds or voice characteristics. We will be especially interested in any particularly bad talker-sentence-vocoder combinations.

7.3 Experimentation with a Third Listening Procedure

A disadvantage of the rank-ordering task described in Section 6 is that it does not directly relate the quality of a given system on one talker or sentence to its quality on another speaker or sentence. Such comparisons must be inferred. A given system could always be given the lowest rank, even though it might be relatively good on one sentence, and relatively poor on another. This fact would not be represented in the data. It is possible, but difficult, to measure relationships of this sort by means of a pair-comparison task; the problem is that the similarity of two recordings is obviously affected by whether or not they were spoken by the same talker.

Therefore, a rating experiment was designed in which listeners assign numbers directly to the quality of all system-talker-sentence combinations. Altogether, there are 504 stimulus sentences to be rated (6 talkers x 6 sentences x 14 "systems"). Since it is well-known that sequential effects can have large influences in tasks of this sort (e.g., Huggins, 1968), we have generated a presentation order in which each talker follows each other talker an equal number of times, and similarly for the sentences. In addition, each system follows each other system an approximately equal number of times (in practice between 2 and 4 times. Furthermore, the order is approximately balanced for serial order. That is, each speaker, sentence and system occur equally often early in the sequence as late in the sequence. As a result of this counter-balancing, we think a single presentation order is probably sufficient, although we may add a second order later.

The judgments required of the listeners will be that of assigning demerits (one mark per demerit) to each sentence they hear. This minimizes the constraints on the subjects to a single anchor. Since each of the 36 speaker-sentence combinations occurs once in its unprocessed form, these natural versions provide an anchor of "perfect" quality, which should, therefore, get zero demerits. To allow for criterion drifts in the early part of the experiment, and to provide a rough check on repeatability, the first 96 stimuli will be repeated at the end of the test, in the same sequence. Criterion drift will be checked by comparing average ratings given to blocks of 12 stimuli on their first and second presentations, and repeatability by comparing individual ratings within the blocks.

7.4 Analysis of the Rank-Order Data that Have Already been Obtained and Those that Will Be Obtained as Described in Paragraphs 7.1 and 7.3 with Two Multidimensional Scaling (MDS) Programs.

The two MDS programs that will be used, INDSCAL and MDPREF were developed at Bell Telephone Laboratories, and have been provided to us for this purpose. Both programs require modifications before being run on BBN TENEX. These modifications have been nearly completed, and we expect to be able to use the programs to analyze our data very soon. The purpose of this analysis is to provide a powerful test for internal consistency of the data. If the "quality spaces" in which the systems are placed on the basis of different types of data are similar to each other, this has two strong implications. First, it means that the experimental measurements are all tapping into the same psychological space, and, second, we can select for future testing whichever of the experimental procedures involves the least effort. If the quality spaces turn out to be similar, this fact can also provide empirical justification for the particular MDS method that generated those spaces.

7.5 Design and Conducting of Phoneme-Specific Tests

If a vocoding system degrades the quality of speech, it is very likely that the degradation is unevenly distributed over the phonemes. If so, some phonemes may be liable to confusion with others, when the powerful redundancies implicit in the syntax and semantics of meaningful speech are removed. The phoneme-specific tests will measure this tendency directly. They measure how well listeners can make such distinctions as voiced-voiceless, stop-fricative, nasal-nonnasal, front vowels vs. back vowels, etc. The procedure involves administering a number of brief nonsense-syllable tests, each of which is designed to look at particular phonetic distinctions, using a closed set of response items.

A complete set of test forms has now been prepared, and we expect to record the tests with two speakers this month. Following

this, the items will be processed, initially by two versions of LPC vocoders to provide stimulus materials for pilot experimentation. These processed recordings will then be presented to listeners and the data will be subjected to a confusion-matrix analysis. On the basis of the results, we will make a decision as to which test types are contributing the most informative data. Further processing of a subset of the tests by a wider range of systems will be carried out.

7.6 Identification of a Set of Descriptors that Might be Used Effectively by Listeners to Characterize Qualitative Aspects of Speech, and Development of a Procedure for Obtaining Listening Characterizations of Speech Samples in Terms of These Descriptors

Examples of such descriptors might be: wheezy, husky, guttural, clear, clicky, nasal, muffled, crisp, muted, warbly, smooth, quavering, etc.

7.7 Acquisition of Speech Samples from LPC Vocoder Systems of the Other ARPA Contractors and Incorporation of This Material in Further Listening Tests

We will attempt to get our own speech material (described in Section 3.3 of this report) processed by the "best" of the systems of the other ARPA contractors, and also one or two of the better channel-vocoders developed over the last few years.

7.8 Experimentation (probably several small studies) Designed to Answer Some Detailed Questions concerning the Effects of Specific LPC Vocoder Parameter Variations on Speech Quality

Examples of questions for which we would like to have an objective answer include the following:

a. For a given average bit rate, how much does one gain in quality by using a variable as opposed to a fixed frame transmission rate?

b. How does going from a fixed to a variable order LPC analysis affect quality?

c. How is quality affected as quantization step size is varied over a large range?

d. Are log-area ratios better than log-error ratios as spectral sensitivity measures, in quantization?

e. Does the use of Rosenberg's shaped excitation pulses produce better quality than unshaped pulses?

f. Does pitch-synchronous analysis and/or resynthesis give better quality than time-synchronous?

g. What is the effect on quality of smoothing of parameters before transmission?

h. Does correcting formant bandwidths before synthesis improve quality?

We should stress that we do not expect to be able to provide answers to all of the above questions. They were enumerated simply to indicate the range of problems we are considering as worthy of study.

REFERENCES

Denes, P. B. On the statistics of spoken English. Journal of the Acoustical Society of America, 1963, 35, 892-904.

Huggins, A.W.F. The perception of timing in natural speech: I. Compensation within the syllable. Language and Speech, 1968, 11, 1-11.

Kalikow, D. N. & Swets, J. A. Experiments with computer-controlled displays in second-language learning. IEEE Transactions on Audio and Electroacoustics, 1972, AU-20, 23-28.

Makhoul, J. I. & Wolf, J. J. Linear prediction and the spectral analysis of speech. BBN Report No. 2304, August 1972.

Nickerson, R. S., Kalikow, D. N., & Stevens, K. N. A computer-based system of speech-training aids for the deaf: A progress report. BBN Report No. 2901, September 1974. Abbreviated version published in AFIPS Conference Proceedings, 1974, 43, 125-126. Also submitted to American Annals of the Deaf.

Nickerson, R. S. & Stevens, K. N. Teaching speech to the deaf: Can a computer help? Proceedings, Association for Computing Machinery, August 1972, 240-252; and IEEE Transactions on Audio and Electro-acoustics, 1973, AU-21, 445-455.

Stevens, K. N., Kalikow, D. N., & Willemain, T. R. The use of a miniature accelerometer for detecting glottal waveforms and nasality. BBN Report No. 2907, September 1974. Accepted for publication, Journal of Speech and Hearing Research.

Stevens, K. N., Nickerson, R. S., Boothroyd, A., & Rollins, A. Assessment of nasality in the speech of deaf children. BBN Report No. 2902, September 1974.

Stevens, K. N., Nickerson, R. S., Rollins, A., & Boothroyd, A. Use of a visual display of nasalization to facilitate training of velar control for deaf speakers. BBN Report No. 2899, September 1974.

Williams, E. J. Experimental designs balanced for pairs of residual effects. Australian Journal of Scientific Research A, 1950, 3, 351.

Appendix A

INSTRUCTIONS FOR VOCODER EVALUATION RANK ORDERING OF QUALITY

This experiment is designed to compare the quality of speech processed through several (14) vocoding systems. Each of 36 sentence tokens (actually 6 different sentences, each spoken by 6 different speakers) has been processed through each of the 14 vocoders, and the processed sentences have then been recorded on Language Master cards. The cards are stored in six boxes, one for each speaker, and within a box there are six groups of cards held together by a rubber band. Each group contains the 14 different processed versions of one sentence. The speakers are identified by a two-letter code (AR, JB, DK, RS, DD, PF), and the 6 sentences are identified by the digits 1 to 6. Each card has written on it the speaker's code and the sentence number, at the top left, and a random identification number between 101 and 604 at the top right.

At the start of the experiment, you will be given a sheet of paper that tells you the order in which you are to listen to the sentences, and 36 answer sheets. Start at the top left corner of the order sheet and work down the columns of the matrix, in order, from left to right.

Procedure Sequence

1. Read the next sentence code from the order sheet. Suppose the next sentence code is PF4.

2. Take the fourth group of cards, labeled PF4, from the box labeled PF, and remove the rubber band.

3. Check that the cards are ordered so that the random identification numbers are in ascending order.

4. Play each card, and put it into the toast rack so that when you have played all 14 cards, the best sounding card is at the front of the rack and the quality gets progressively worse from front to back. You will probably get the order only approximately right the first time, so play any cards or groups of cards you wish to hear again, in any order you like, as often as you like. Finally, play through the whole set, from best to worst, to check that you are satisfied with your ordering. Remember, there are NO CORRECT ANSWERS. It is YOUR JUDGMENT of the order that we want.

5. When you are satisfied with the order, copy the random identification numbers from the top right corner of each card onto the answer sheet, being careful to copy the numbers correctly, and not to change the order of the cards. Also fill out the other data asked for at the top of the answer sheet.

6. Perform any other ratings you are asked to make, and fill in the sheets appropriately.

7. Put the 14 cards back in their original order, so that the random identification numbers are in ascending order. Replace the rubber band, and put the group of cards back in the right place in the right box.

8. Return to Step 1.

Further Comments

You may do more than one column of the order matrix in one day, but, if you do, take a break of at least an hour after finishing a column. Please be gentle with the cards. If treated roughly the tape can separate from the card. Use the toast rack for sorting the cards, both to rank order them and to restore the serial order.